Institut de recherche et documentation en économie de la santé

Données déclaratives versus données médico-administratives dans l'identification des pathologies et des situations de handicap

Bibliographie thématique

Juin 2026

Centre de documentation de l'Irdes

Véronique Suhard

Synthèses & Bibliographies

ISSN 2606-0272 **ISBN : 978-2-87812-665-5**

Table des matières

Problématique 3

Validité, limites et comparabilité des données déclaratives et données administratives 4

Appariements entre les bases de données médico-administratives et les données d'enquêtes déclaratives 39

Problématique

L'utilisation des données en santé publique repose principalement sur deux grandes sources : les données médico-administratives, issues des systèmes de soins, et les données déclaratives, collectées par enquêtes auprès des individus. Ces deux types de données offrent des informations complémentaires sur l'état de santé des populations, mais présentent également des limites spécifiques. Les données administratives sont généralement exhaustives et précises pour les événements médicaux objectivables, tandis que les données déclaratives permettent de saisir des dimensions subjectives, telles que la morbidité ressentie ou certains comportements de santé.

Cependant, de nombreuses études mettent en évidence des divergences importantes entre ces sources, notamment en termes d'estimation de la prévalence des maladies ou de l'utilisation des soins. Ces écarts s'expliquent par différents biais, comme les erreurs de déclaration, le biais de rappel ou les limites des systèmes de codage.

Une première bibliographie[1] avait été réalisée en 2019 et avait ciblé la problématique du repérage des maladies et du handicap dans les bases médico-administratives, notamment via les algorithmes d'identification des cas.

Cette nouvelle bibliographie, plus orientée sur les aspects méthodologiques, la complète en s'intéressant d'une part, à la littérature, examinant la validité, les différences et la complémentarité entre données administratives et auto-déclarées ; elle recense, d'autre part, quelques études sur les méthodes d'appariement de ces deux sources qui permettent d'améliorer leur utilisation en santé publique.

[1] Safon, M.-O. (2019). Le repérage des maladies et du handicap dans les bases médico-administratives. Paris, Irdes

Validité, limites et comparabilité des données déclaratives et données administratives

Alper, H. E., Brite, J., Cone, J. E., et al. (2021). "Comparison of prevalence and exposure-disease associations using self-report and hospitalization data among enrollees of the world trade center health registry." BMC Med Res Methodol **21**(1): 162.

> BACKGROUND: Although many studies have investigated agreement between survey and hospitalization data for disease prevalence, it is unknown whether exposure-chronic disease associations vary based on data collection method. We investigated agreement between self-report and administrative data for the following: 1) disease prevalence, and 2) the accuracy of self-reported hospitalization in the last 12 months, and 3) the association of seven chronic diseases (rheumatoid arthritis, hypertension, heart attack, stroke, asthma, diabetes, hyperlipidemia) with four measures of 9/11 exposure. METHODS: Enrollees of the World Trade Center Health Registry who resided in New York State were included (N = 18,206). Hospitalization data for chronic diseases were obtained from the New York State Planning and Research Cooperative System (SPARCS). Prevalence for each disease and concordance measures (kappa, sensitivity, specificity, positive agreement, and negative agreement) were calculated. In addition, the associations of the seven chronic diseases with the four measures of exposure were evaluated using logistic regression. RESULTS: Self-report disease prevalence ranged from moderately high (40.5% for hyperlipidemia) to low (3.8% for heart attack). Self-report prevalence was at least twice that obtained from administrative data for all seven chronic diseases. Kappa ranged from 0.35 (stroke) to 0.04 (rheumatoid arthritis). Self-reported hospitalizations within the last 12 months showed little overlap with actual hospitalization data. Agreement for exposure-disease associations was good over the twenty-eight exposure-disease pairs studied. CONCLUSIONS: Agreement was good for exposure-disease associations, modest for disease prevalence, and poor for self-reported hospitalizations. Neither self-report nor administrative data can be treated as the "gold standard." Which source to use depends on the availability and context of data, and the disease under study.

Banham, D., Hawthorne, G., Goldney, R., et al. (2014). "Health-related quality of life (HRQoL) changes in South Australia: comparison of burden of disease morbidity and survey-based health utility estimates." Health Qual Life Outcomes **12**: 113.

> BACKGROUND: Global research shows a clear transition in health outcomes over the past two decades where improved survival was accompanied by lower health related quality of life (HRQoL) as measured by morbidity and disability. These trends suggest the need to better understand changes in population HRQoL. This paper compares two perspectives on population HRQoL change using burden of disease morbidity estimates from administrative data and self-reports from random and representative population surveys. METHODS: South Australian administrative data including inpatient hospital activity, cancer and communicable disease registrations were used within a Burden of Disease study framework to quantify morbidity as Prevalent Years of Life lived with Disease and injury related illness (PYLD) for 1999 to 2008. Self-reported HRQoL was measured using the Assessment of Quality of Life (AQoL) in face to face interviews with at least 3000 respondents in each of South Australia's Health Omnibus Surveys (HOS) in 1998, 2004 and 2008. RESULTS: Age specific PYLD rates for those aged 75 or more increased by 5.1%. HRQoL dis-utility in this age group also increased significantly and beyond the minimally important difference threshold. Underlying increased dis-utility were greater difficulties in independent living (particularly requiring help with household tasks) and psychological well-being (as influenced by pain, discomfort and difficulty sleeping). CONCLUSIONS: Consistent with increased quantity of life being accompanied by reduced HRQoL, the analysis indicates older people in South Australia experienced increased morbidity in the decade to 2008. The results warrant routine monitoring of health disutility at a population level and improvement to the supply and scope of administrative data.

Berete, F., Demarest, S., Charafeddine, R., et al. (2020). "Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium." Arch Public Health **78**(1): 120.

Background Health administrative data were increasingly used for chronic diseases (CDs) surveillance purposes. This cross sectional study explored the agreement between Belgian compulsory health insurance (BCHI) data and Belgian health interview survey (BHIS) data for asserting CDs. Methods Individual BHIS 2013 data were linked with BCHI data using the unique national register number. The study population included all participants of the BHIS 2013 aged 15 years and older. Linkage was possible for 93% of BHIS-participants, resulting in a study sample of 8474 individuals. For seven CDs disease status was available both through self-reported information from the BHIS and algorithms based on ATC-codes of disease-specific medication, developed on demand of the National Institute for Health and Disability Insurance (NIHDI). CD prevalence rates from both data sources were compared. Agreement was measured using sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) assuming BHIS data as gold standard. Kappa statistic was also calculated. Participants' sociodemographic and health status characteristics associated with agreement were tested using logistic regression for each CD. Results Prevalence from BCHI data was significantly higher for CVDs but significantly lower for COPD and asthma. No significant difference was found between the two data sources for the remaining CDs. Sensitivity was 83% for CVDs, 78% for diabetes and ranged from 27 to 67% for the other CDs. Specificity was excellent for all CDs (above 98%) except for CVDs. The highest PPV was found for Parkinson's disease (83%) and ranged from 41 to 75% for the remaining CDs. Irrespective of the CDs, the NPV was excellent. Kappa statistic was good for diabetes, CVDs, Parkinson's disease and thyroid disorders, moderate for epilepsy and fair for COPD and asthma. Agreement between BHIS and BCHI data is affected by individual sociodemographic characteristics and health status, although these effects varied across CDs. Conclusions NHIDI's CDs case definitions are an acceptable alternative to identify cases of diabetes, CVDs, Parkinson's disease and thyroid disorders but yield in a significant underestimated number of patients suffering from asthma and COPD. Further research is needed to refine the definitions of CDs from administrative data.

Berete, F., Van der Heyden, J., Demarest, S., et al. (2021). "Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias." Eur J Public Health **31**(1): 214-220.

Background: The validity of self-reported mammography uptake is often questioned. We assessed the related selection and reporting biases among women aged 50-69years in the Belgian Health Interview Survey (BHIS) using reimbursement data for mammography stemming from the Belgian Compulsory Health Insurance organizations (BCHI). Methods: Individual BHIS 2013 data (n = 1040) were linked to BCHI data 2010-13 (BHIS-BCHI sample). Being reimbursed for mammography within the last 2-years was used as the gold standard. Selection bias was assessed by comparing BHIS estimates reimbursement rates in BHIS-BCHI with similar estimates from the Echantillon Permanent/Permanente Steekproef (EPS), a random sample of BCHI data, while reporting bias was investigated by comparing self-reported versus reimbursement information in the BHIS-BCHI. Reporting bias was further explored through measures of agreement and logistic regression. Results: Mammography uptake rates based on self-reported information and reimbursement from the BHIS-BCHI were 75.5% and 69.8%, respectively. In the EPS, it was 64.1%. The validity is significantly affected by both selection bias {relative size =8.93% [95% confidence interval (CI): 3.21-14.64]} and reporting bias [relative size =8.22% (95% CI: 0.76-15.68)]. Sensitivity was excellent (93.7%), while the specificity was fair (66.4%). The agreement was moderate (kappa =0.63). Women born in non-EU countries (OR= 2.81, 95% CI: 1.54-5.13), with high household income (OR= 1.27, 95% CI: 1.02-1.60) and those reporting poor perceived health (OR= 1.41, 95% CI: 1.14-1.73) were more likely to inaccurately report their mammography uptake. Conclusions: The validity of self-reported mammography uptake in women aged 50-69 years is affected by both selection and reporting bias. Both administrative and survey data are complementary when assessing mammography uptake.

Bhandari, A. et Wagner, T. H. (2006). "Self-Reported Utilization of Health Care Services: Improving Measurement and Accuracy." Medical Care Research and Review **63**: 217 - 235.

Self-report is often used to estimate health care utilization. However, the accuracy of such data is of paramount concern. The authors conducted a systematic review of 42 studies that evaluated the accuracy of self-report utilization data, where utilization was defined as a visit to a clinical provider or entity. They also present a broad conceptual model that identifies major issues to consider when collecting, analyzing, and reporting such data. The results show that self-report data are of variable accuracy. Factors that affect accuracy include (1) sample population and cognitive abilities, (2) recall time frame, (3) type of utilization, (4) utilization frequency, (5) questionnaire design, (6) mode of data collection, and (7) memory aids and probes.

Bhaskar, R., Noon, J. et O'Hara, B. J. (2019). "The Errors in Reporting Medicare Coverage: A Comparison of Survey Data and Administrative Records." J Aging Health **31**(10): 1806-1829.

Objectives: We examine survey reporting of Medicare coverage of the older population by evaluating discordance between survey responses and administrative records. Method: We link data from the 2014 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and 2014 Medicare Enrollment Database to evaluate the extent to which individuals misreport Medicare coverage in the CPS ASEC. Using regression analyses, we assess factors associated with misreporting. Results: We find the CPS ASEC undercounts the population aged 65 years and older with Medicare by 4.5%. Misreporting of Medicare coverage is associated with citizenship status, immigration year of entry, employment, coverage of other household members, and imputation of Medicare responses. Adjusting for misreporting, Medicare coverage among older individuals increases from 93.4% to 95.6%. Discussion: The CPS ASEC underestimates Medicare coverage for the older population. Administrative records may be useful to evaluate and improve survey imputation of Medicare coverage when missing.

Blais, C., Rochette, L., Hamel, D., et al. (2014). "Prevalence, incidence, awareness and control of hypertension in the province of Quebec: perspective from administrative and survey data." Can J Public Health **105**(1): e79-85.

OBJECTIVES: Hypertension is a major risk factor for cardiovascular diseases. Nearly one adult in four was diagnosed with hypertension in 2007-2008 in Canada. One of the objectives of this study was to determine whether the prevalence of hypertension in Quebec as assessed using administrative data is comparable to that specifically measured, especially for the elderly population. METHODS: Trends in prevalence, incidence and mortality were examined using the Quebec Integrated Chronic Disease Surveillance System built from grouping numerous administrative databases from 1996-1997 to 2009-2010. Blood pressure measurements, hypertension prevalence, awareness and control were obtained in 1,706 Quebecers in the combined cycles of the Canadian Health Measures Survey. RESULTS: Using administrative databases, 23.6% [95% confidence interval, 23.5-23.6] of the Quebec population (n=1,433,400) aged ≥20 years was diagnosed with hypertension in 2009-2010, an increase of 32.1% compared to 2000 2001. The incidence decreased by 27.3%. Among people aged ≥65 years, the prevalence rose to 69.0% [95% CI: 68.8 69.2] in women and 61.7% [95% CI: 61.5-61.9] in men. For people aged 20-79 years, the prevalence of hypertension was lower with the administrative data compared to the survey (20.2% and 23.1%, respectively). The level of awareness, treatment and control were 84.3%, 83.1% and 67.9%, respectively. CONCLUSION: The prevalence of hypertension derived from administrative data is comparable to that obtained with a health measured survey. Elderly women (≥65 years) are a very high-risk subgroup. The levels of awareness, treatment and control of hypertension in Quebec are very high.

Boscardin, C. K., Gonzales, R., Bradley, K. L., et al. (2015). "Predicting cost of care using self-reported health status data." BMC Health Serv Res **15**: 406.

Background: We examined whether self-reported employee health status data can improve the performance of administrative data-based models for predicting future high health costs, and develop a predictive model for predicting new high cost individuals. Methods: This retrospective cohort study used data from 8,917 Safeway employees self-insured by Safeway during 2008 and 2009. We created models using step-wise multivariable logistic regression starting with health services use data, then

socio-demographic data, and finally adding the self-reported health status data to the model. Results: Adding self-reported health data to the baseline model that included only administrative data (health services use and demographic variables; c-statistic = 0.63) increased the model" predictive power (c-statistic = 0.70). Risk factors associated with being a new high cost individual in 2009 were: 1) had one or more ED visits in 2008 (adjusted OR: 1.87, 95 % CI: 1.52, 2.30), 2) had one or more hospitalizations in 2008 (adjusted OR: 1.95, 95 % CI: 1.38, 2.77), 3) being female (adjusted OR: 1.34, 95 % CI: 1.16, 1.55), 4) increasing age (compared with age 18-35, adjusted OR for 36-49 years: 1.28; 95 % CI: 1.03, 1.60; adjusted OR for 50-64 years: 1.92, 95 % CI: 1.55, 2.39; adjusted OR for 65+ years: 3.75, 95 % CI: 2.67, 2.23), 5) the presence of self-reported depression (adjusted OR: 1.53, 95 % CI: 1.29, 1.81), 6) chronic pain (adjusted OR: 2.22, 95 % CI: 1.81, 2.72), 7) diabetes (adjusted OR: 1.73, 95 % CI: 1.35, 2.23), 8) high blood pressure (adjusted OR: 1.42, 95 % CI: 1.21, 1.67), and 9) above average BMI (adjusted OR: 1.20, 95 % CI: 1.04, 1.38).Discussion: The comparison of the models between the full sample and the sample without the previous high cost members indicated significant differences in the predictors. This has important implications for models using only the health service use (administrative data) given that the past high cost is significantly correlated with future high cost and often drive the predictive models. Conclusions: Self-reported health data improved the ability of our model to identify individuals at risk for being high cost beyond what was possible with administrative data alone.

Buajitti, E., Chiodo, S. et Rosella, L. C. (2020). "Agreement between area- and individual-level income measures in a population-based cohort: Implications for population health research." SSM-Population Health **10**.

Socioeconomic status is an important determinant of health, the measurement of which is of great significance to population health research. However, individual-level socioeconomic factors are absent from much health administrative data, resulting in widespread use of area-level measures in their place. This study aims to clarify the role of individual- and area-level socioeconomic status in Ontario, Canada, through comparison of income measures. Using data from four cycles (2005-2012) of the Canadian Community Health Survey, we assessed concordance between individual- and area-level income quintiles using percent agreement and Kappa statistics. Individual-level characteristics were compared at baseline. Cumulative adult premature mortality was calculated for 5-years following interview. Rates were calculated separately for area-level and individual-level income, and jointly for each combination of income groups. Multivariable negative binomial models were fit to estimate associations between area- and individual-level income quintile and premature mortality after adjustment for basic demographics (age, sex, interview cycle) and key risk factors (alcohol, smoking, physical activity, and body mass index). Agreement between individual- and area-level income measures was low. Kappa statistics for same and similar (i.e +/- 1 quintile) measures were 0.11 and 0.48, indicating low and moderate agreement, respectively. Socioeconomic disparities in premature mortality were greater for individual-level income than area-level income. When rates were stratified by both area- and individual-level income quintiles simultaneously, individual-level income gradients persisted within each area-level income group. The association between income and premature mortality was significant for both measures, including after full adjustment. Area-level socioeconomic status is an inappropriate proxy for missing individual-level data. The low agreement between area- and individual-level income measures and differences in demographic profile indicate that the two socioeconomic status measures do not capture the same population groups. However, our findings demonstrate that both individual- and area-level income measures are associated with premature mortality, and describe unique socioeconomic inequities.

Bulloch, A. G., Currie, S., Guyn, L., et al. (2011). "Estimates of the treated prevalence of bipolar disorders by mental health services in the general population: comparison of results from administrative and health survey data." Chronic Dis Inj Can **31**(3): 129-134.

INTRODUCTION: Informed provision of population mental health services requires accurate estimates of disease burden. METHODS: We estimated the treated prevalence of bipolar disorders by mental health services in the Calgary Zone, a catchment area in Alberta with a population of over one million. Administrative data in a central repository provides information of mental health care contacts for about 95% of publically funded mental health services. We compared this treated prevalence against self-reported data in the 2002 Canadian Community Health Survey: Mental Health and Well-Being

(CCHS 1.2). RESULTS: Of the 63 016 individuals aged 18 years plus treated in the Calgary Zone in 2002-2008, 3659 (5.81%) and 1065 (1.70%) were diagnosed with bipolar I and bipolar II disorder, respectively. The estimated treated population prevalence of these disorders was 0.41% and 0.12%, respectively. We estimated that 0.44% to 1.17% of the Canadian population was being treated by psychiatrists for bipolar I disorder from CCHS 1.2. DISCUSSION: For bipolar I disorder the estimate based on local administrative data is close to the lower end of the health survey range. The degree of agreement in our estimates reinforces the utility of administrative data repositories in the surveillance of chronic mental disorders.

Corbellini, C., Andreoni, B., Ansaloni, L., et al. (2018). "Reliability and validity assessment of administrative databases in measuring the quality of rectal cancer management." Tumori **104**(1): 51-59.

PURPOSE: Measurement and monitoring of the quality of care using a core set of quality measures are increasing in health service research. Although administrative databases include limited clinical data, they offer an attractive source for quality measurement. The purpose of this study, therefore, was to evaluate the completeness of different administrative data sources compared to a clinical survey in evaluating rectal cancer cases. METHODS: Between May 2012 and November 2014, a clinical survey was done on 498 Lombardy patients who had rectal cancer and underwent surgical resection. These collected data were compared with the information extracted from administrative sources including Hospital Discharge Dataset, drug database, daycare activity data, fee-exemption database, and regional screening program database. The agreement evaluation was performed using a set of 12 quality indicators. RESULTS: Patient complexity was a difficult indicator to measure for lack of clinical data. Preoperative staging was another suboptimal indicator due to the frequent missing administrative registration of tests performed. The agreement between the 2 data sources regarding chemoradiotherapy treatments was high. Screening detection, minimally invasive techniques, length of stay, and unpreventable readmissions were detected as reliable quality indicators. Postoperative morbidity could be a useful indicator but its agreement was lower, as expected. CONCLUSIONS: Healthcare administrative databases are large and real-time collected repositories of data useful in measuring quality in a healthcare system. Our investigation reveals that the reliability of indicators varies between them. Ideally, a combination of data from both sources could be used in order to improve usefulness of less reliable indicators.

Coste, J., Mandereau-Bruno, L., Constantinou, P., et al. (2025). "Healthcare claims and health interview survey data for chronic disease surveillance: agreement and comparative validity of prevalence indicators for 20 chronic conditions in a general population sample in France." Eur J Public Health **35**(4): 624-634.

Healthcare claims data are increasingly used to derive chronic condition (CC) surveillance indicators, although comparative evidence with self-reported data remains scarce. We explored the agreement and comparative validity (concurrent and predictive) of 20 CC prevalence indicators independently constructed using the French National Health Data System (SNDS) and Health, Health Care, and Insurance Survey (ESPS 2010-2014). Individual data from 5039 ESPS participants aged >= 25 years, representative of the 2010 French general population, were linked to the SNDS. Follow-up data included a 2014 health self-assessment and 5-year mortality. We considered 20 CCs with corresponding SNDS case-identifying algorithms and self-reported information from ESPS, including most cardiovascular diseases and frequent cancers. Kappa statistics assessed agreement between CC indicators across databases. Polytomous and dichotomous logistic regression assessed determinants of disagreement between sources and associations of indicators with health outcomes (concurrent and predictive validity). Prevalence values were much higher with survey data except for hypertension, diabetes, thyroid disorders, epilepsy, and most cancers for which they were closer (+/- 20%) to claims data. Agreement between CC indicators varied from the strongest (hypertension, diabetes, thyroid disorders, most cancers) to the weakest (cardiac rhythm disorders, peptic ulcer, chronic liver diseases). Sex, age, and multimorbidity strongly influenced agreement. Most claims database indicators were more strongly associated with health outcomes. Health interview surveys and healthcare claims-derived indicators are not interchangeable given their specific determinants. Since no general rule applies to all CCs, the advantages and disadvantages of each data source should be closely considered in case-to-case analysis.

Darvishian, M., Chu, J., Simkin, J., et al. (2022). "Agreement between self-report and administrative health data on occurrence of non-cancer chronic disease among participants of the BC generations project." Front Epidemiol **2**: 1054485.

Population-based studies of non-cancer chronic disease often rely on self-reported data for disease diagnosis, which may be incomplete, unreliable and suffer from bias. Recently, the British Columbia Generations Project (BCGP; n = 29,736) linked self-reported chronic disease history data to a Chronic Disease Registry (CDR) that applied algorithms to administrative health data to ascertain diagnoses of multiple chronic diseases in the Province of British Columbia. For the 10 diseases captured by both self-report and the CDR, including asthma, chronic obstructive pulmonary disease (COPD), diabetes, hypertension, multiple sclerosis, myocardial infarction, osteoarthritis, osteoporosis, rheumatoid arthritis, and stroke, we calculated Cohen's kappa coefficient to examine concordance of chronic disease status (i.e., ever/never diagnosed) between the data sources. Using CDR data as the gold standard, we also calculated sensitivity, specificity, and positive-predictive value (PPV) for self-reported chronic disease occurrence. The prevalence of each chronic disease was similar across both data sources. Substantial levels of concordance (0.66-0.73) and moderate to high sensitivities (0.64-0.92), specificities (0.98-0.99) and PPVs (0.55-0.84) were observed for diabetes, hypertension, multiple sclerosis, and myocardial infarction. We did observe degree of concordance to vary by age, sex, body mass index (BMI), health perception, and ethnicity across most of the chronic diseases that were evaluated. While administrative health data are imperfect, they are less likely to suffer from bias, making them a reasonable gold standard. Our results demonstrate that for at least some chronic diseases, self-report may be a reasonable method for case ascertainment. However, characteristics of the study population will likely have impacts on the quality of the data.

Dohouin, I., Laberge, M., Lacasse, A., et al. (2024). "Identification of Mood Disorders in Self-Reported Versus Health Administrative Data." Brain Behav **14**(11): e70126.

BACKGROUND: Producing relevant knowledge on the prevalence of mood disorders (MDs) requires a clear identification of people living with the condition. Analyzing this multifaceted disease from the perspective of health administrative data and population-based surveys could contribute to document inconsistencies between these data sources and highlight the strengths and limitations of each methodological approaches. OBJECTIVES: The aim of this study was to estimate the prevalence of MD disease, assess concordance of MD patterns in population-based surveys versus health administrative data, and investigate statistical differences in characteristics between individuals presenting the disease in each data sources. METHODS: This study used the Care Trajectories-Enriched Data (TorSaDE) cohort. The TorSaDE cohort is built by merging five waves of the Canadian Community Health Survey (CCHS) with health administrative data of the province of Quebec, Canada. The sample includes individuals who participated in at least one round of CCHS and for whom evidence of use of health services in the year of CCHS completion and the year before were present in health administrative data. The cohort was split into four groups based on the presence and absence of MD in self-reported versus health administrative data. Groups' characteristics were compared using chi-square tests and ANOVA. RESULTS: The study cohort was composed of 96,079 individuals, of which 10,418 (10.8%) had MD, regardless of the data sources. Self-reported prevalence of MD was 6.03%, while the prevalence from health administrative data was about 7.79%. Estimates showed a low level of concordance between the two measures, as only 27.4% of people presenting this medical condition were identified in both data sources. Furthermore, individuals identified with MD only in survey data had poorer socioeconomic outcomes but better health outcomes than those from the concordant group (i.e., identified in both data sources). In addition, people presenting MD in health administrative data only had better socioeconomic and health outcomes than those who reported MD diagnosis only in survey data. CONCLUSION: Findings suggest that each measure capture different specific subpopulations. Estimates obtained from each source should thus be contextualized and interpreted with caution.

Doiron, D., Fiebig, D. G., Johar, M., et al. (2014). "Does self-assessed health measure health?" Applied Economics **47**(2): 180-194.

Despite concerns about reporting biases and interpretation, self-assessed health (SAH) remains the measure of health most used by researchers, in part reflecting its ease of collection and in part the observed correlation between SAH and objective measures of health. Using a unique Australian data set, which consists of survey data linked to administrative individual medical records, we present empirical evidence demonstrating that SAH indeed predicts future health, as measured by hospitalizations, out-of-hospital medical services and prescription drugs. Our large sample size allows very disaggregate analysis and we find that SAH predicts more serious, chronic illnesses better than less serious illnesses. Finally, we compare the predictive power of SAH relative to administrative data and an extensive set of self-reported health measures; SAH does not add to the predictive power of future utilization when the administrative data is included and improves prediction only marginally when the extensive survey-based health measures are included. Clearly there is value in the more extensive survey and administrative health data as well as greater cost of collection.

Donnenfeld, M., Espagnacq, M., Regaert, C., et al. (2025). "Limitations motrices et organiques : confrontation de l'enquête santé européenne et de l'algorithme RISH sur les données médico-administratives du SNDS." J Epidemiol Popul Health **73 Suppl 1**: 202885.

Drapeau, A., Boyer, R. et Diallo, F. B. (2011). "Discrepancies between survey and administrative data on the use of mental health services in the general population: findings from a study conducted in Quebec." BMC Public Health **11**.

Background: Population surveys and health services registers are the main source of data for the management of public health. Yet, the validity of survey data on the use of mental health services has been questioned repeatedly due to the sensitive nature of mental illness and to the risk of recall bias. The main objectives of this study were to compare data on the use of mental health services from a large scale population survey and a national health services register and to identify the factors associated with the discrepancies observed between these two sources of data. Methods: This study was based on the individual linkage of data from the cycle 1.2 of the Canadian Community Health Survey (CCHS-1.2) and from the health services register of the Regie de l'assurance maladie du Quebec (RAMQ). The RAMQ is the governmental agency managing the Quebec national health insurance program. The analyses mostly focused on the 637 Quebecer respondents who were recorded as users of mental health services in the RAMQ and who were self-reported users or non users of these services in the CCHS-1.2. Results: Roughly 75%, of those recorded as users of mental health services users in the RAMQ's register did not report using mental health services in the CCHS-1.2. The odds of disagreement between survey and administrative data were higher in seniors, individuals with a lower level of education, legal or de facto spouses and mothers of young children. They were lower in individuals with a psychiatric disorder and in frequent and more recent users of mental health services according to the RAMQ's register. Conclusions: These findings support the hypotheses that social desirability and recall bias are likely to affect the self-reported use of mental health services in a population survey. They stress the need to refine the investigation of mental health services in population surveys and to combine survey and administrative data, whenever possible, to obtain an optimal estimation of the population need for mental health care.

Dufour, I., Vedel, I. et Quesnel-Vallée, A. (2022). "Identification of Major Cognitive Disorders in Self-Reported versus Administrative Health Data: A Cohort Study in Quebec." Journal of Alzheimer's Disease **89**: 1091 - 1101.

Duncan, L., Georgiades, K., Wang, L., et al. (2022). "Estimating prevalence of child and youth mental disorder and mental health-related service contacts: a comparison of survey data and linked administrative health data." Epidemiol Psychiatr Sci **31**: e35.

AIMS: Prevalence estimates of child and youth mental disorder and mental health-related service contacts are needed for policy formulation, research, advocacy and resource allocation. Our aim is to compare prevalence estimates of child and youth mental disorder and mental health-related service

contacts derived from general population survey data v. linked administrative health data. METHODS: Provincially representative 2014 Ontario Child Health Study data were linked to administrative health records for 5563 children and youth aged 4-17 in Ontario. Emotional disorders (mood and anxiety) and attention-deficit/hyperactivity disorder were assessed using a standardised diagnostic interview in the survey and using diagnostic codes in administrative health data. Physician-based mental health-related service contacts were assessed using parent self-reports from the survey and administrative data related to mental health-related diagnostic codes. Prevalence estimates were calculated and compared based on one-sample z-tests and ratios of survey data to administrative data-based prevalence. Sensitivity, specificity and agreement between classifications were compared using kappa. Prevalence estimates were calculated by age, sex and geography sub-groups and consistent group differences across data source were counted. RESULTS: Disorder prevalence and service contact estimates were significantly higher in survey data in all cases, except for mood disorder. Ratios of survey data to administrative data-based prevalence varied, ranging from 0.80 (mood) to 11.01 (attention-deficit/hyperactivity disorder). Specificity was high (0.98-1.00), sensitivity was low (0.07-0.41) and agreement ranged from slight (kappa = 0.13) to moderate (kappa = 0.46). Out of 18 sub-group difference comparisons, half were non-significant in either data source. In the remaining nine comparisons, the only significant differences between groups that were consistent across data source were for sex-based differences (attention-deficit/hyperactivity disorder and service contacts). There were no consistent age- or geography-based differences in prevalence across data sources. CONCLUSIONS: Our findings suggest that conclusions drawn about prevalence, service contacts and sub-group differences in these estimates are dependent on data source. Further research is needed to understand who and what is being captured by each source. Researchers should conduct data linkage where possible to access and compare multiple sources of information.

Edwards, J., Thind, A., Stranges, S., et al. (2020). "Concordance between health administrative data and survey-derived diagnoses for mood and anxiety disorders." Acta Psychiatr Scand **141**(4): 385-395.

OBJECTIVE: To assess whether estimates of survey structured interview diagnoses of mood and anxiety disorders were concordant with diagnoses of these disorders obtained from health administrative data. METHODS: All Ontario respondents to the 2012 Canadian Community Health Survey-Mental Health (CCHS-MH) were linked to health administrative databases at ICES (formerly known as the Institute for Clinical Evaluative Sciences). Survey structured interview diagnoses were compared with health administrative data diagnoses obtained using a standardized algorithm. We used modified Poisson regression analyses to assess whether socio-demographic factors were associated with concordance between the two measures. RESULTS: Of the 4157 Ontarians included in our sample, 20.4% had either a structured interview diagnosis (13.9%) or health administrative diagnosis (10.4%) of a mood or anxiety disorder. There was high discordance between measures, with only 19.4% agreement. Migrant status, age, employment, and income were associated with discordance between measures. CONCLUSIONS: Our findings indicate that previous estimates of the 12-month prevalence of mood and anxiety disorders in Ontario may be underestimating the true prevalence, and that population-based surveys and health administrative data may be capturing different groups of people. Understanding the limitations of data commonly used in epidemiologic studies is a key foundation for improving population-based estimates of mental disorders.

Emerson, S. D., McLinden, T., Sereda, P., et al. (2024). "Secondary use of routinely collected administrative health data for epidemiologic research: Answering research questions using data collected for a different purpose." International Journal of Population Data Science (IJPDS) **9**(1): 1-12.

The use of routinely collected administrative health data for research can provide unique insights to inform decision-making and, ultimately, support better public health outcomes. Yet, since these data are primarily collected to administer healthcare service delivery, challenges exist when using such data for secondary purposes, namely epidemiologic research. Many of these challenges stem from the researcher's lack of control over the quality and consistency of data collection, and - furthermore - a lessened understanding of the data being analyzed. That said, we assert that these challenges can be partly mitigated through careful, systematic use of these data in epidemiologic research. This article presents considerations derived from experiences analyzing administrative health data (e.g.,

healthcare practitioner billings, hospitalizations, and prescription medication data) in the Canadian province of British Columbia (population of over 5 million in 2024), though we believe the underlying principles generalize beyond this region. Key considerations were organized around four themes: 1) Know the data and their primary use (understand their scope and limitations); 2) Understand classification and coding systems (appreciate the nuances regarding classification systems, versions, how they are employed in the primary uses of the data, and querying the values); 3) Transform data into meaningful forms (process data and apply identification algorithms, when necessary); 4) Recognize the importance of validity when defining analytic variables (make meaningful inferences based on data/algorithms). Although this article is not an exhaustive list of all considerations, we believe that it will provide pragmatic insights for those interested in leveraging administrative health data for epidemiologic research.

Fortin, M., Haggerty, J., Sanche, S., et al. (2017). "Self-reported versus health administrative data: implications for assessing chronic illness burden in populations. A cross-sectional study." CMAJ Open **5**(3): E729-e733.

BACKGROUND: Various data sources may be used to document the presence of chronic medical conditions. This study examined the agreement between self-reported and health administrative data. METHODS: A randomly selected cohort of participants aged 25-75 years recruited by telephone from the general population of Quebec reported on the presence of 1 or more chronic conditions from a candidate list of 12 conditions: diabetes, hypertension, thyroid disorder, any cardiac disease, cancer diagnosis in the previous 5 years (including melanoma but excluding other skin cancers), asthma, osteoarthritis, rheumatoid arthritis or lupus, osteoporosis, chronic obstructive pulmonary disease, intestinal disease and hypercholesterolemia. We also used health administrative data from Quebec's universal health insurance provider to identify participants' chronic conditions. Unique identifiers allowed linkage of both data sources to the individual participant. The frequencies of the 12 conditions and the prevalence of multimorbidity (≥ 2, ≥ 3 and ≥ 4 conditions) were analyzed for each data source. RESULTS: We analyzed data for 1177 participants (mean age 53 [standard deviation 12.4] yr; 684 women [58.1%]). We found low (but varied) agreement between the 2 data sources, with the poorest agreement for hypercholesterolemia (κ = 0.04 [95% confidence interval (CI) 0.01 to 0.07]) and the best for diabetes (κ = 0.82 [95% CI 0.76 to 0.88]). Prevalence estimates of multimorbidity obtained with health administrative data were lower than those obtained with self-reported data regardless of the operational definition used. Most participants with multimorbidity were identified by self-report. INTERPRETATION: We argue for the use of self-reported chronic conditions in the study of multimorbidity, as health administrative data based on the billing system in Quebec seem to underestimate the prevalence of many chronic conditions, which results in biased estimates of multimorbidity.

Frank, J. (2016). "Comparing nationwide prevalences of hypertension and depression based on claims data and survey data: An example from Germany." Health Policy **120**(9): 1061-1069.

INTRODUCTION: Coded diagnoses in claims data offer a comprehensive basis for health sciences and health policy decisions. For example, morbidity-based risk adjustment schemes use coded diagnoses to allocate resources. Therefore a routinely performed validation is important. Data reconciliation with medical records would be first best, but is not possible here. This paper validates population-based prevalences of hypertension and depression based on claims data by comparing them with prevalences stem from two different epidemiological survey data. METHOD: Data sources accessible are a nationwide sample based on outpatient claims data (GSPR), a nationwide health interview and examination survey (DEGS1) and a nationwide telephone interview survey (GEDA). The analysis includes SHI-insured aged 18-79 who live in 2010 in Germany. RESULTS: There was high agreement for hypertension prevalences between GSPR (28.98% [28.95-29.02]) and DEGS1 (28.0% [26.5-29.5]) but not with GEDA (22.9% [22.1-23.7]). The agreement for depression prevalences was high between the two surveys (DEGS1: 7.6% [6.7-8.5] and GEDA: 6.7% [6.3-7.2]) and moderate compared to GSPR (12.23% [12.21-12.26]). CONCLUSION: For an objectifiable disease, such as hypertension, diagnostic coding with claims data seems to be valid to be used for risk adjustment in German outpatient health care. Even though depression prevalences differ between claims data and survey data, more effort is required to understand the magnitude of a reference systems impact on prevalence estimates.

Gavrielov-Yusim, N. et Friger, M. (2014). "Use of administrative medical databases in population-based research." J Epidemiol Community Health **68**(3): 283-287.

Administrative medical databases are massive repositories of data collected in healthcare for various purposes. Such databases are maintained in hospitals, health maintenance organisations and health insurance organisations. Administrative databases may contain medical claims for reimbursement, records of health services, medical procedures, prescriptions, and diagnoses information. It is clear that such systems may provide a valuable variety of clinical and demographic information as well as an on-going process of data collection. In general, information gathering in these databases does not initially presume and is not planned for research purposes. Nonetheless, administrative databases may be used as a robust research tool. In this article, we address the subject of public health research that employs administrative data. We discuss the biases and the limitations of such research, as well as other important epidemiological and biostatistical key points specific to administrative database studies.

Gindi, R. et Cohen, R. A. (2012). "Assessing measurement error in Medicare coverage from the National Health Interview Survey." Medicare Medicaid Res Rev **2**(2).

OBJECTIVES: Using linked administrative data, to validate Medicare coverage estimates among adults aged 65 or older from the National Health Interview Survey (NHIS), and to assess the impact of a recently added Medicare probe question on the validity of these estimates. DATA SOURCES: Linked 2005 NHIS and Master Beneficiary Record and Payment History Update System files from the Social Security Administration (SSA). STUDY DESIGN: We compared Medicare coverage reported on NHIS with "benchmark" benefit records from SSA. PRINCIPAL FINDINGS: With the addition of the probe question, more reports of coverage were captured, and the agreement between the NHIS-reported coverage and SSA records increased from 88% to 95%. Few additional overreports were observed. CONCLUSIONS: Increased accuracy of the Medicare coverage status of NHIS participants was achieved with the Medicare probe question. Though some misclassification remains, data users interested in Medicare coverage as an outcome or correlate can use this survey measure with confidence.

Gini, R., Francesconi, P., Mazzaglia, G., et al. (2013). "Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey." BMC Public Health **13**: 15.

BACKGROUND: Administrative databases are widely available and have been extensively used to provide estimates of chronic disease prevalence for the purpose of surveillance of both geographical and temporal trends. There are, however, other sources of data available, such as medical records from primary care and national surveys. In this paper we compare disease prevalence estimates obtained from these three different data sources. METHODS: Data from general practitioners (GP) and administrative transactions for health services were collected from five Italian regions (Veneto, Emilia Romagna, Tuscany, Marche and Sicily) belonging to all the three macroareas of the country (North, Center, South). Crude prevalence estimates were calculated by data source and region for diabetes, ischaemic heart disease, heart failure and chronic obstructive pulmonary disease (COPD). For diabetes and COPD, prevalence estimates were also obtained from a national health survey. When necessary, estimates were adjusted for completeness of data ascertainment. RESULTS: Crude prevalence estimates of diabetes in administrative databases (range: from 4.8% to 7.1%) were lower than corresponding GP (6.2%-8.5%) and survey-based estimates (5.1%-7.5%). Geographical trends were similar in the three sources and estimates based on treatment were the same, while estimates adjusted for completeness of ascertainment (6.1%-8.8%) were slightly higher. For ischaemic heart disease administrative and GP data sources were fairly consistent, with prevalence ranging from 3.7% to 4.7% and from 3.3% to 4.9%, respectively. In the case of heart failure administrative estimates were consistently higher than GPs' estimates in all five regions, the highest difference being 1.4% vs 1.1%. For COPD the estimates from administrative data, ranging from 3.1% to 5.2%, fell into the confidence interval of the Survey estimates in four regions, but failed to detect the higher prevalence in the most Southern region (4.0% in administrative data vs 6.8% in survey data). The prevalence estimates for

COPD from GP data were consistently higher than the corresponding estimates from the other two sources. CONCLUSION: This study supports the use of data from Italian administrative databases to estimate geographic differences in population prevalence of ischaemic heart disease, treated diabetes, diabetes mellitus and heart failure. The algorithm for COPD used in this study requires further refinement.

Goldberg, M., Quantin, C., Gueguen, A., et al. (2008). "Bases de données médico-administratives et épidémiologie : intérêts et limites." Courrier des Statistiques(124): 59-70.

Les épidémiologistes cherchent à décrire les phénomènes de santé des populations, à comprendre l'histoire naturelle des maladies et à juger, en termes de causalité, du rôle de facteurs de risque sur la santé. Dans tous ces domaines, les difficultés méthodologiques sont difficiles à contrôler. Elles proviennent essentiellement des biais potentiellement induits par divers phénomènes de sélection qui interviennent lors de l'inclusion des sujets dans l'étude épidémiologique et durant le suivi de celle-ci lorsqu'elle est longitudinale. Certaines solutions efficaces pour prendre en compte la non-participation et pour le calcul des pondérations impliquent le recours à des bases de données nationales réputées exhaustives (bases de données de santé et socioprofessionnelles essentiellement). Les problèmes rencontrés, aussi bien méthodologiques que ceux liés à des questions d'éthique et de confidentialité, sont illustrés par l'exemple de la cohorte Constances en cours de mise en place ; la plate-forme scientifique et technique « Plastico », actuellement en phase de préfiguration, peut apporter une aide considérable à la mise en œuvre des solutions proposées.

Golden, G., Wang, L. et Reid, G. J. (2025). "Child and youth chronic physical health conditions: a comparison of survey data and linked administrative health data in Ontario." BMC Pediatr **25**(1): 741.

BACKGROUND: Population-based studies in Canada and the United States estimate chronic physical health conditions affect between 20 to 30% of children aged 0 to 17. Challenges in measuring chronic conditions include the use of inconsistent definitions and algorithms that capture a limited number of conditions. Thus, we developed a chronic health condition (CHC) algorithm using administrative data to determine whether a child has a CHC based on (1) the diagnosis recorded for the visit, (2) the number of visits, and (3) within a specific reference period. METHODS: Data were from the cross-sectional 2014 Ontario Child Health Study, linked with Ontario Health Insurance Plan (OHIP) administrative health data. Unweighted prevalence estimates and agreement analyses (Cohen's Kappa, sensitivity, specificity) were used to compare the survey parent-reported and algorithm-based presence of a CHC. RESULTS: 31.8% and 27.1% of children and youth had a CHC based on administrative and survey data, respectively. Agreement between administrative and survey data was poor (k = 0.17). Among a few specific conditions, agreement varied depending on the type of condition (e.g., diabetes k = 0.77 vs health conditions k = 0.21). CONCLUSION: We found considerable discrepancies between administrative and survey-reported data. The results highlight the importance of using algorithms developed from multiple datasets to examine complex research questions, such as the measurement of chronicity.

Gontijo Guerra, S., Berbiche, D. et Vasiliadis, H. M. (2019). "Measuring multimorbidity in older adults: comparing different data sources." BMC Geriatr **19**(1): 166.

BACKGROUND: Multimorbidity is a global health issue, particularly for older adults in the primary care setting. An adequate portrayal of its epidemiology is essential to properly identify and understand the health care needs of this population. This study aimed to compare the differences in the prevalence of selected chronic conditions and multimorbidity, including its associated characteristics, using health survey/self-reported (SR) information only, administrative (Adm) data only and the combined (either) sources. METHODS: This was a secondary analysis of survey data from the first cycle of the Longitudinal Survey on Senior's Health and Health Services Use linked to health-Adm data. The analytical sample consisted of 1625 community-dwelling older adults (>/=65 years) recruited in the waiting rooms of primary health clinics in a selected administrative region of the province of Quebec. Seventeen chronic conditions were assessed according to two different data sources. We examined the differences in the observed prevalence of chronic conditions and multimorbidity and the

agreement between data sources. RESULTS: The prevalence of each of the 17 chronic conditions ranged from 1.2 to 68.7% depending on the data source. The agreement between different data sources was highly variable, with kappa coefficients (kappa) ranging from 0.05 to 0.73. Multimorbidity was very high in this population, with an estimated prevalence of up to 95.9%. In addition, we found that the association between sociodemographic and behavioural factors and the presence of multimorbidity varied according to the different data sources and thresholds. CONCLUSIONS: This is the first study to simultaneously investigate chronic conditions and multimorbidity prevalence among primary care older adults using combined SR and health-Adm data. Our results call attention to (1) the possibility of underestimating cases when using a single data source and (2) the potential benefits of integrating information from different data sources to increase case identification. This is an important aspect of characterizing the health care needs of this fast-growing population.

Gravely, A. A., Cutting, A., Nugent, S., et al. (2011). "Validity of PTSD diagnoses in VA administrative data: Comparison of VA administrative PTSD diagnoses to self-reported PTSD Checklist scores." J Rehabil Res Dev **48**(1): 21-30.

Little research has been done on the validity of posttraumatic stress disorder (PTSD) diagnoses that are found in Department of Veterans Affairs (VA) administrative data, even though they are often used in VA research. We compared PTSD diagnoses found in VA administrative data with PTSD Checklist (PCL) scores self-reported by 4,777 newly diagnosed participants in a national postal survey study. Using PCL scores of at least 50 as the gold standard, we compared positive predictive values (PPVs) for at least one versus at least two PTSD diagnoses (found within 4 months of the first) in VA administrative data overall and by subgroups of interest: age, sex, and clinic where first diagnosed. The overall PPV was 75% for at least one PTSD diagnosis and 82% for at least two PTSD diagnoses. Similarly, the PPV significantly increased for all subgroup analyses when at least two PTSD diagnoses were used. The increase in PPV was greatest for those first diagnosed in primary care and for those older than 65. To select a sample of veterans with more definitive PTSD from administrative data, researchers should select those veterans with at least two PTSD diagnoses as opposed to at least one.

Griffith, L. E., Gruneir, A., Fisher, K. A., et al. (2021). "The impact of multimorbidity level and functional limitations on the accuracy of using self-reported survey data compared to administrative data to measure general practitioner and specialist visits in community-living adults." BMC Health Serv Res **21**(1): 1123.

BACKGROUND: Researchers often use survey data to study the effect of health and social variables on physician use, but how self-reported physician use compares to administrative data, the gold standard, in particular within the context of multimorbidity and functional limitations remains unclear. We examine whether multimorbidity and functional limitations are related to agreement between self-reported and administrative data for physician use. METHODS: Cross-sectional data from 52,854 Ontario participants of the Canadian Community Health Survey linked to administrative data were used to assess agreement on physician use. The number of general practitioner (GP) and specialist visits in the previous year was assessed using both data sources; multimorbidity and functional limitation were from self-report. RESULTS: Fewer participants self-reported GP visits (84.8%) compared to administrative data (89.1%), but more self-reported specialist visits (69.2% vs. 64.9%). Sensitivity was higher for GP visits (≥90% for all multimorbidity levels) compared to specialist visits (approximately 75% for 0 to 90% for 4+ chronic conditions). Specificity started higher for GP than specialist visits but decreased more swiftly with multimorbidity level; in both cases, specificity levels fell below 50%. Functional limitations, age and sex did not impact the patterns of sensitivity and specificity seen across level of multimorbidity. CONCLUSIONS: Countries around the world collect health surveys to inform health policy and planning, but the extent to which these are linked with administrative, or similar, data are limited. Our study illustrates the potential for misclassification of physician use in self-report data and the need for sensitivity analyses or other corrections.

Griffith, L. E., Gruneir, A., Fisher, K. A., et al. (2020). "Measuring multimorbidity series-an overlooked complexity comparison of self-report vs. administrative data in community-living adults: paper 2. Prevalence estimates depend on the data source." J Clin Epidemiol **124**: 163-172.

OBJECTIVE: The objective of the study was to compare multimorbidity prevalence using self-reported and administrative data and identify factors associated with agreement between data sources. STUDY DESIGN AND SETTING: Self-reported cross-sectional data from four Canadian Community Health Survey waves were linked to administrative data in Ontario, Canada. Multimorbidity prevalence was examined using two definitions, 2+ and 3+ chronic conditions (CCs). Agreement between data sources was assessed using Kappa and Phi statistics. Logistic regression was used to estimate associations between agreement and sociodemographic, health behavior, and health status variables for each multimorbidity definition. RESULTS: Regardless of multimorbidity definition, prevalence was higher using administrative data (2+ CCs: 55.5% vs. 47.1%; 3+ CCs: 30.0% vs. 24.2%). Agreement between data sources was moderate (2+ CCs K = 0.482; 3+ CCs K = 0.442), and while associated with sociodemographic, health behavior, and health status factors, the magnitude and sometimes direction of association differed by multimorbidity definition. CONCLUSION: A better understanding is needed of what factors influence individuals' reporting of CCs and how they align with what is in administrative data as policy makers need a solid evidence base on which to make decisions for health planning. Our results suggest that data sources may need to be triangulated to provide accurate estimates of multimorbidity for health services planning and policy.

Griffith, L. E., Gruneir, A., Fisher, K. A., et al. (2020). "The hidden complexity of measuring number of chronic conditions using administrative and self-report data: A short report." Journal of Comorbidity **10**: 2235042X20931287.

OBJECTIVE: To examine agreement between administrative and self-reported data on the number of and constituent chronic conditions (CCs) used to measure multimorbidity. STUDY DESIGN AND SETTING: Cross-sectional self-reported survey data from four Canadian Community Health Survey waves were linked to administrative data for residents of Ontario, Canada. Agreement for each of 12 CCs was assessed using kappa (kappa) statistics. For the overall number of CCs, perfect agreement was defined as agreement on both the number and constituent CCs. Jackknife methods were used to assess the impact of individual CCs on perfect agreement. RESULTS: The level of chance-adjusted agreement between self-report and administrative data for individual CCs varied widely, from kappa = 5.5% (inflammatory bowel disease) to kappa = 77.5% (diabetes), and there was no clear pattern on whether using administrative data or self-reported data led to higher prevalence estimates. Only 26.9% of participants had perfect agreement on the number and constituent CCs; 10.6% agreed on the number but not constituent CCs. The impact of each CC on perfect agreement depended on both the level of agreement and the prevalence of the individual CC. CONCLUSION: Our results show that measuring agreement on multimorbidity is more complex than for individual CCs and that even small levels of individual condition disagreement can have a large impact on the agreement on the number of CCs.

Groen, J. A. (2012). "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." Journal of Official Statistics **28**(2): 173-198.

With increased attention to administrative data for statistical purposes, analyses of the quality of administrative data and comparisons to survey data are greatly needed. This article presents a methodology for identifying sources of error in administrative and survey data and for identifying sources of differences between administrative and survey estimates. The first part of the methodology is a statistical decomposition of the difference between administrative and survey estimates. The second part investigates the causes of measurement error and the factors associated with differences between what the same respondents report in the survey and the administrative records. I illustrate this methodology using a case study of the monthly employment figures gathered from U.S. businesses. This analysis demonstrates that both administrative data and survey data may contain errors and that reporting procedures are likely to differ between the two types of data. The article also identifies practical ways to assess data quality.

Gruneir, A., Griffith, L. E., Fisher, K., et al. (2020). "Measuring multimorbidity series. An overlooked complexity - Comparison of self-report vs. administrative data in community-living adults: Paper 3. Agreement across data sources and implications for estimating associations with health service use." J Clin Epidemiol **124**: 173-182.

Objective The objective of this study is to describe agreement between administrative and self-report data on the number and type of chronic conditions (CCs) and determine whether associations between CC count and health service use differ by data source.

Guerard, B., Omachonu, V., Harvey, R. A., et al. (2016). "The Influence of Respondent Characteristics on the Validity of Self-Reported Survey Responses." Health Services Research **51**(3): 937-952.

Objective. To examine concordance between member self-reports and the organization's administrative claims data for two key health factors: number of chronic conditions, and number of prescription drugs. Data. Medicare Advantage plan claims data and member survey data from 2011 to 2012. Design. Mailed surveys to 15,000 members, enrolled minimum 6 months, drawn from a random sample of primary care physician practices with at least 200 members. Methods. Descriptive statistics were generated for extent of concordance. Multivariable logistic regressions were used to analyze the association of selected respondent characteristics with likelihood of concordance. Findings. Concordance for number of chronic conditions was 58.4 percent, with 27.3 percent under-reporting, 14.2 percent over-reporting. Concordance for number of prescription drugs was 56.6 percent with 38.9 percent under-reporting, 4.5 percent over-reporting. Number of prescriptions and assistance in survey completion were associated with higher likelihood of concordance for chronic conditions. Assistance in survey completion and number of chronic conditions were associated with higher concordance, and age and number of prescriptions were associated with lower concordance, for prescription drugs. Conclusions. Self-reported number of chronic conditions and prescription medications are not in high concordance with claims data. Health care researchers and policy makers using patient self-reported data should be aware of these potential biases.

Hale, M. D., Santorelli, G., Brundle, C., et al. (2019). "A cross-sectional study assessing agreement between self-reported and general practice-recorded health conditions among community dwelling older adults." Age Ageing **49**(1): 135-140.

BACKGROUND: self-reported data regarding health conditions are utilised in both clinical practice and research, but their agreement with general practice records is variable. The extent of this variability is poorly studied amongst older adults, particularly amongst those with multiple health conditions, cognitive impairment or frailty. This study investigates the agreement between self-reported and general practice-recorded data amongst such patients and the impact of participant factors on this agreement. METHODS: data on health conditions was collected from participants in the Community Ageing Research 75+ (CARE75+) study (n = 964) by self-report during face-to-face assessment and interrogation of the participants' general practice electronic health records. Agreement between self-report and practice records was assessed using Kappa statistics and the effect of participant demographics using logistic regression. RESULTS: agreement ranged from K = 0.25 to 1.00. The presence of ≥2 health conditions modified agreement for cancer (odds ratio, OR:0.62, 95%confidence interval, CI:0.42-0.94), diabetes (OR:0.55, 95%CI:0.38-0.80), dementia (OR:2.82, 95%CI:1.31-6.13) and visual impairment (OR:3.85, 95%CI:1.71-8.62). Frailty reduced agreement for cerebrovascular disease (OR:0.45, 95%CI:0.23-0.89), heart failure (OR:0.40, 95%CI:0.19-0.84) and rheumatoid arthritis (OR:0.41, 95%CI:0.23-0.75). Cognitive impairment reduced agreement for dementia (OR:0.36, 95%CI:0.21-0.62), diabetes (OR:0.47, 95%CI:0.33-0.67), heart failure (OR:0.53, 95%CI:0.35-0.80), visual impairment (OR:0.42, 95%CI:0.25-0.69) and rheumatoid arthritis (OR:0.53, 95%CI:0.37-0.76). CONCLUSIONS: significant variability exists for agreement between self-reported and general practice-recorded comorbidities. This is further affected by an individual's health conditions. This study is the first to assess frailty as a factor modifying agreement and highlights the importance of utilising the general practice records as the gold standard for data collection from older adults.

Harron, K., Dibben, C., Boyd, J., et al. (2017). "Challenges in administrative data linkage for research." Big Data Soc **4**(2): 2053951717745678.

Linkage of population-based administrative data is a valuable tool for combining detailed individual-level information from different sources for research. While not a substitute for classical studies based on primary data collection, analyses of linked administrative data can answer questions that require

large sample sizes or detailed data on hard-to-reach populations, and generate evidence with a high level of external validity and applicability for policy making. There are unique challenges in the appropriate research use of linked administrative data, for example with respect to bias from linkage errors where records cannot be linked or are linked together incorrectly. For confidentiality and other reasons, the separation of data linkage processes and analysis of linked data is generally regarded as best practice. However, the 'black box' of data linkage can make it difficult for researchers to judge the reliability of the resulting linked data for their required purposes. This article aims to provide an overview of challenges in linking administrative data for research. We aim to increase understanding of the implications of (i) the data linkage environment and privacy preservation; (ii) the linkage process itself (including data preparation, and deterministic and probabilistic linkage methods) and (iii) linkage quality and potential bias in linked data. We draw on examples from a number of countries to illustrate a range of approaches for data linkage in different contexts.

Hoffmann, J., Haastert, B., Brüne, M., et al. (2018). "How do patients with diabetes report their comorbidities? Comparison with administrative data." Clinical Epidemiology **10**: 499-509.

Aims: Patients with diabetes are probably often unaware of their comorbidities. We estimated agreement between self-reported comorbidities and administrative data. Methods: In a random sample of 464 diabetes patients, data from a questionnaire asking about the presence of 14 comorbidities closely related to diabetes were individually linked with statutory health insurance data. Results: Specificities were >97%, except cardiac insufficiency (94.5%), eye diseases (93.8%), peripheral arterial disease (92.6%), hypertension (90.9%), and peripheral neuropathy (85.8%). Sensitivities were <60%, except amputation (100%), hypertension (83.1%), and myocardial infarction (67.2%). A few positive predictive values were >90% (hypertension, myocardial infarction, and eye disease), and six were below 70%. Six negative predictive values were >90%, and two <70% (hypertension and eye disease). Total agreement was between 42.7% (eye disease) and 100% (dialysis and amputation). Overall, substantial agreement was observed for three morbidities (kappa 0.61-0.80: hypertension, myocardial infarction, and amputation). Moderate agreement (kappa 0.41-0.60) was estimated for angina pectoris, heart failure, stroke, peripheral neuropathy, and kidney disease. Factors associated with agreement were the number of comorbidities, diabetes duration, age, sex, and education. Conclusions: Myocardial infarction and amputation were well reported by patients as comorbidities; eye diseases and foot ulceration rather poorly, particularly in older, male, or less educated patients. Patient information needs improving.

Hure, A. J., Chojenta, C. L., Powers, J. R., et al. (2015). "Validity and reliability of stillbirth data using linked self-reported and administrative datasets." J Epidemiol **25**(1): 30-37.

BACKGROUND: A high rate of stillbirth was previously observed in the Australian Longitudinal Study of Women's Health (ALSWH). Our primary objective was to test the validity and reliability of self-reported stillbirth data linked to state-based administrative datasets. METHODS: Self-reported data, collected as part of the ALSWH cohort born in 1973-1978, were linked to three administrative datasets for women in New South Wales, Australia (n = 4374): the Midwives Data Collection; Admitted Patient Data Collection; and Perinatal Death Review Database. Linkages were obtained from the Centre for Health Record Linkage for the period 1996-2009. True cases of stillbirth were defined by being consistently recorded in two or more independent data sources. Sensitivity, specificity, positive predictive value, negative predictive value, percent agreement, and kappa statistics were calculated for each dataset. RESULTS: Forty-nine women reported 53 stillbirths. No dataset was 100% accurate. The administrative datasets performed better than self-reported data, with high accuracy and agreement. Self-reported data showed high sensitivity (100%) but low specificity (30%), meaning women who had a stillbirth always reported it, but there was also over-reporting of stillbirths. About half of the misreported cases in the ALSWH were able to be removed by identifying inconsistencies in longitudinal data. CONCLUSIONS: Data linkage provides great opportunity to assess the validity and reliability of self-reported study data. Conversely, self-reported study data can help to resolve inconsistencies in administrative datasets. Quantifying the strengths and limitations of both self-reported and administrative data can improve epidemiological research, especially by guiding methods and interpretation of findings.

Hyde, J. S., Harrati, A. et Center for Retirement Research, B. C. (2021). "The Alignment Between Self-Reported and Administrative Measures of Application to and Receipt of Federal Disability Benefits in the Health and Retirement Study." SSRN Electronic Journal.

Jutte, D. P., Roos, L. L. et Brownell, M. D. (2011). "Administrative record linkage as a tool for public health research." Annu Rev Public Health **32**: 91-108.

Linked administrative databases offer a powerful resource for studying important public health issues. Methods developed and implemented in several jurisdictions across the globe have achieved high-quality linkages for conducting health and social research without compromising confidentiality. Key data available for linkage include health services utilization, population registries, place of residence, family ties, educational outcomes, and use of social services. Linking events for large populations of individuals across disparate sources and over time permits a range of research possibilities, including the capacity to study low-prevalence exposure-disease associations, multiple outcome domains within the same cohort of individuals, service utilization and chronic disease patterns, and life course and transgenerational transmission of health. Limited information on variables such as individual-level socioeconomic status (SES) and social supports is outweighed by strengths that include comprehensive follow-up, continuous data collection, objective measures, and relatively low expense. Ever advancing methodologies and data holdings guarantee that research using linked administrative databases will make increasingly important contributions to public health research.

Kemp, A., Preen, D. B., Saunders, C., et al. (2013). "Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia." BMC Medical Research Methodology **13**.

Background: Statutory State-based cancer registries are considered the 'gold standard' for researchers identifying cancer cases in Australia, but research using self-report or administrative health datasets (e. g. hospital records) may not have linkage to a Cancer Registry and need to identify cases. This study investigated the validity of administrative and self-reported data compared with records in a State-wide Cancer Registry in identifying invasive breast cancer cases. Methods: Cases of invasive breast cancer recorded on the New South Wales (NSW) Cancer Registry between July 2004 and December 2008 (the study period) were identified for women in the 45 and Up Study. Registry cases were separately compared with suspected cases ascertained from: i) administrative hospital separations records; ii) outpatient medical service claims; iii) prescription medicines claims; and iv) the 45 and Up Study baseline survey. Ascertainment flags included diagnosis codes, surgeries (e. g. lumpectomy), services (e. g. radiotherapy), and medicines used for breast cancer, as well as self-reported diagnosis. Positive predictive value (PPV), sensitivity and specificity were calculated for flags within individual datasets, and for combinations of flags across multiple datasets. Results: Of 143,010 women in the 45 and Up Study, 2039 (1.4%) had an invasive breast tumour recorded on the NSW Cancer Registry during the study period. All of the breast cancer flags examined had high specificity (>97.5%). Of the flags from individual datasets, hospital-derived 'lumpectomy and diagnosis of invasive breast cancer' and '(lumpectomyor mastectomy) and diagnosis of invasive breast cancer' had the greatest PPV (89% and 88%, respectively); the later having greater sensitivity (59% and 82%, respectively). The flag with the highest sensitivity and PPV >= 85% was ' diagnosis of invasive breast cancer' (both 86%). Self-reported breast cancer diagnosis had a PPV of 50% and sensitivity of 85%, and breast radiotherapy had a PPV of 73% and a sensitivity of 58% compared with Cancer Registry records. The combination of flags with the greatest PPV and sensitivity was '(lumpectomy or mastectomy) and (diagnosis of invasive breast cancer or breast radiotherapy)' (PPV and sensitivity 83%). Conclusions: In the absence of Cancer Registry data, administrative and self-reported data can be used to accurately identify cases of invasive breast cancer for sample identification, removing cases from a sample, or risk adjustment. Invasive breast cancer can be accurately identified using hospital-derived diagnosis alone or in combination with surgeries and breast radiotherapy.

Kilkenny, M. F., Dalli, L. L., Sanders, A., et al. (2024). "Comparison of comorbidities of stroke collected in administrative data, surveys, clinical trials and cohort studies." Health Inf Manag **53**(2): 104-111.

BACKGROUND: Administrative data are used extensively for research purposes, but there remains limited information on the quality of these data for identifying comorbidities related to stroke. OBJECTIVE: To compare the prevalence of comorbidities of stroke identified using International Classification Diseases, Australian Modification (ICD-10-AM) or Anatomical Therapeutic Chemical codes, with those from (i) self-reported data and (ii) published studies. METHOD: The cohort included patients with stroke or transient ischaemic attack admitted to hospitals (2012-2016; Victoria and Queensland) in the Australian Stroke Clinical Registry (N = 26,111). Data were linked with hospital and pharmaceutical datasets to ascertain comorbidities using published algorithms. The sensitivity, specificity, and positive predictive value of these comorbidities were compared with survey responses from 623 patients (reference standard). An indirect comparison was also performed with clinical data from published stroke studies. RESULTS: The sensitivity of hospital ICD-10-AM data was poor for most comorbidities, except for diabetes (93.0%). Specificity was excellent for all comorbidities (87-96%), except for hypertension (70.5%). Compared to published stroke studies (3 clinical trials and 1 incidence study), the prevalence of diabetes and atrial fibrillation in our cohort was similar using ICD-10-AM codes, but lower for dyslipidaemia and anxiety/depression. Whereas in the pharmaceutical dispensing data, the sensitivity was excellent for dyslipidaemia (94%) and modest for anxiety/depression (77%). In the pharmaceutical data, specificity was modest for hypertension (78%) and anxiety or depression (76%), but specificity was poor for dyslipidaemia (19%) and heart disease (46%). CONCLUSION: Variation was observed in the reporting of comorbidities of stroke in administrative data, and consideration of multiple sources of data may be necessary for research. Further work is needed to improve coding and clinical documentation for reporting of comorbidities in administrative data.

Leong, A., Dasgupta, K., Bernatsky, S., et al. (2013). "Systematic Review and Meta-Analysis of Validation Studies on a Diabetes Case Definition from Health Administrative Records." PLoS One **8**(10).

Objectives: Health administrative data are frequently used for diabetes surveillance. We aimed to determine the sensitivity and specificity of a commonly-used diabetes case definition (two physician claims or one hospital discharge abstract record within a two-year period) and their potential effect on prevalence estimation. Methods: Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, we searched Medline (from 1950) and Embase (from 1980) databases for validation studies through August 2012 (keywords: "diabetes mellitus''; "administrative databases''; "validation studies''). Reviewers abstracted data with standardized forms and assessed quality using Quality Assessment of Diagnostic Accuracy Studies (QUADAS) criteria. A generalized linear model approach to random-effects bivariate regression meta-analysis was used to pool sensitivity and specificity estimates. We applied correction factors derived from pooled sensitivity and specificity estimates to prevalence estimates from national surveillance reports and projected prevalence estimates over 10 years (to 2018). Results: The search strategy identified 1423 abstracts among which 11 studies were deemed relevant and reviewed; 6 of these reported sensitivity and specificity allowing pooling in a meta-analysis. Compared to surveys or medical records, sensitivity was 82.3% (95% CI 75.8, 87.4) and specificity was 97.9% (95% CI 96.5, 98.8). The diabetes case definition underestimated prevalence when it was <= 10.6% and overestimated prevalence otherwise. Conclusion: The diabetes case definition examined misses up to one fifth of diabetes cases and wrongly identifies diabetes in approximately 2% of the population. This may be sufficiently sensitive and specific for surveillance purposes, in particular monitoring prevalence trends. Applying correction factors to adjust prevalence estimates from this definition may be helpful to increase accuracy of estimates.

Leong, A., Dasgupta, K., Chiasson, J. L., et al. (2013). "Estimating the Population Prevalence of Diagnosed and Undiagnosed Diabetes." Diabetes Care **36**(10): 3002-3008.

OBJECTIVE Health administrative data are frequently used for diabetes surveillance, but validation studies are limited, and undiagnosed diabetes has not been considered in previous studies. We compared the test properties of an administrative definition with self-reported diabetes and

estimated prevalence of undiagnosed diabetes by measuring glucose levels in mailed-in capillary blood samples. RESEARCH DESIGN AND METHODSA stratified random sample of 6,247 individuals (Quebec province) was surveyed by telephone and asked to mail in fasting blood samples on filter paper to a central laboratory. An administrative definition was applied (two physician claims or one hospitalization for diabetes within a 2-year period) and compared with self-reported diabetes alone and with self-reported diabetes or elevated blood glucose level (7 mmol/L). Population-level prevalence was estimated with the use of the administrative definition corrected for its sensitivity and specificity. RESULTS Compared with self-reported diabetes, sensitivity and specificity were 84.3% (95% CI 79.3-88.5%) and 97.9% (97.4-98.4%), respectively. Compared with diabetes by self-report and/or glucose testing, sensitivity was lower at 58.2% (52.2-64.6%), whereas specificity was similar at 98.7% (98.0-99.3%). Adjusted for sampling weights, population-level prevalence of physician-diagnosed diabetes was 7.2% (6.3-8.0%). Prevalence of total diabetes (physician-diagnosed and undiagnosed) was 13.4% (11.7-15.0%), indicating that approximate to 40% of diabetes cases are undiagnosed. CONCLUSIONS A substantial proportion of diabetes cases are missed by surveillance methods that use health administrative databases. This finding is concerning because individuals with undiagnosed diabetes are likely to have a delay in treatment and, thus, a higher risk for diabetes-related complications.

Lix, L. M., Yogendran, M. S., Shaw, S. Y., et al. (2010). "Comparing administrative and survey data for ascertaining cases of irritable bowel syndrome: a population-based investigation." BMC Health Serv Res **10**: 31.

BACKGROUND: Administrative and survey data are two key data sources for population-based research about chronic disease. The objectives of this methodological paper are to: (1) estimate agreement between the two data sources for irritable bowel syndrome (IBS) and compare the results to those for inflammatory bowel disease (IBD); (2) compare the frequency of IBS-related diagnoses in administrative data for survey respondents with and without self-reported IBS, and (3) estimate IBS prevalence from both sources. METHODS: This retrospective cohort study used linked administrative and health survey data for 5,134 adults from the province of Manitoba, Canada. Diagnoses in hospital and physician administrative data were investigated for respondents with self-reported IBS, IBD, and no bowel disorder. Agreement between survey and administrative data was estimated using the kappa statistic. The chi2 statistic tested the association between the frequency of IBS-related diagnoses and self-reported IBS. Crude, sex-specific, and age-specific IBS prevalence estimates were calculated from both sources. RESULTS: Overall, 3.0% of the cohort had self-reported IBS, 0.8% had self-reported IBD, and 95.3% reported no bowel disorder. Agreement was poor to fair for IBS and substantially higher for IBD. The most frequent IBS-related diagnoses among the cohort were anxiety disorders (34.4%), symptoms of the abdomen and pelvis (26.9%), and diverticulitis of the intestine (10.6%). Crude IBS prevalence estimates from both sources were lower than those reported previously. CONCLUSIONS: Poor agreement between administrative and survey data for IBS may account for differences in the results of health services and outcomes research using these sources. Further research is needed to identify the optimal method(s) to ascertain IBS cases in both data sources.

Longobardi, T., Walker, J. R., Graff, L. A., et al. (2011). "Health service utilization in IBD: comparison of self-report and administrative data." BMC Health Serv Res **11**.

Background: The reliability of self-report regarding health care utilization in inflammatory bowel disease (IBD) is unknown. If proven reliable, it could help justify self-report as a means of determining health care utilization and associated costs. Methods: The Manitoba IBD Cohort Study is a population-based longitudinal study of participants diagnosed within 7 years of enrollment. Health care utilization was assessed through standardized interview. Participants (n = 352) reported the total number of nights hospitalized, frequency of physician contacts in the prior 12 months and whether the medical contacts were for IBD-related reasons or not. Reports of recent antibiotic use were also recorded. Actual utilization was drawn from the administrative database of Manitoba Health, the single comprehensive provincial health insurer. Results: According to the administrative data, 15% of respondents had an overnight hospitalization, while 10% had an IBD-related hospitalization. Self-report concordance was highly sensitive (92%; 82%) and specific (96%; 97%, respectively). 97% of participants had contact with a physician in the previous year, and 69% had IBD-related visits.

Physician visits were significantly under-reported and there was a trend to over-report the number of nights in hospital. Conclusions: Self-report data can be helpful in evaluating health service utilization, provided that the researcher is aware of the systematic sources of bias. Outpatient visits are well identified by self-report. The discordance for the type of outpatient visit may be either a weakness of self-report or a flaw in diagnosis coding of the administrative data. If administrative data are not available, self-report information may be a cost-effective alternative, particularly for hospitalizations.

Loreau, J.-M., Mayet, A., Tchakamian, S., et al. (2026). "Use of the 'historical' National Health Data System to study infectious diseases." J Epidemiol Popul Health **74**(2): 203393.

Introduction Identifying infectious diseases through French National Health Data System (SNDS), a medico-administrative database, presents specific challenges due to their intrinsic characteristics, such as curability or highly variable prevalence. This study assesses the capacity of the SNDS to accurately identify pathologies and to ensure follow-up of affected individuals. Methods The study focuses on "Certain infectious and parasitic diseases" (ICD-10 Chapters A and B). Eligible conditions were required to correspond to a three-character ICD-10 sub-chapter accounting for fewer than 10,000 discharge summaries (MCO) within the Programme for Medicalisation of Information Systems (PMSI) between 2006 and 2024. It evaluates the influence of different indicators, such as the extraction sources, the quality of record linkage using the national registration number (NIR) and the follow-up of subjects. Results A total of 155,419 individuals across 77 selected ICD-10 sub-chapters were included. On average, 6% of individuals were non-linkable. Subjects identification was primarily achieved through the beneficiary registry (87%). By sub-chapter, the PMSI was the main source of identification, with an average of 95%. The average proportion of individuals still present in the SNDS five years after the initial occurrence was 58%. Discussion These findings are closely linked to the core concept of the SNDS, which relies on health insurance data collection. Data linkage is fundamentally tied to insurance affiliation; consequently, while high-prevalence tropical or sub-tropical diseases are identifiable, longitudinal tracking is more challenging to achieve.

Lujic, S., Simpson, J. M., Zwar, N., et al. (2017). "Multimorbidity in Australia: Comparing estimates derived using administrative data sources and survey data." PLoS One **12**(8).

Background Estimating multimorbidity (presence of two or more chronic conditions) using administrative data is becoming increasingly common. We investigated (1) the concordance of identification of chronic conditions and multimorbidity using self- report survey and administrative datasets; (2) characteristics of people with multimorbidity ascertained using different data sources; and (3) whether the same individuals are classified as multimorbid using different data sources. Methods Baseline survey data for 90,352 participants of the 45 and Up Study D a cohort study of residents of New South Wales, Australia, aged 45 years and over D were linked to prior twoyear pharmaceutical claims and hospital admission records. Concordance of eight selfreport chronic conditions (reference) with claims and hospital data were examined using sensitivity (Sn), positive predictive value (PPV), and kappa (k). The characteristics of people classified as multimorbid were compared using logistic regression modelling. Results Agreement was found to be highest for diabetes in both hospital and claims data (k=0.79, 0.78; Sn = 79%, 72%; PPV = 86%, 90%). The prevalence of multimorbidity was highest using self-report data (37.4%), followed by claims data (36.1%) and hospital data (19.3%). Combining all three datasets identified a total of 46 683 (52%) people with multimorbidity, with half of these identified using a single dataset only, and up to 20% identified on all three datasets. Characteristics of persons with and without multimorbidity were generally similar. However, the age gradient was more pronounced and people speaking a language other than English at home were more likely to be identified as multimorbid by administrative data. Conclusions Different individuals, with different combinations of conditions, are identified as multimorbid when different data sources are used. As such, caution should be applied when ascertaining morbidity from a single data source as the agreement between self-report and administrative data is generally poor. Future multimorbidity research exploring specific disease combinations and clusters of diseases that commonly co-occur, rather than a simple disease count, is likely to provide more useful insights into the complex care needs of individuals with multiple chronic conditions.

Lujic, S., Watson, D. E., Randall, D. A., et al. (2014). "Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia." BMJ OPEN **4**(9): e005768.

OBJECTIVES: To investigate the nature and potential implications of under-reporting of morbidity information in administrative hospital data. SETTING AND PARTICIPANTS: Retrospective analysis of linked self-report and administrative hospital data for 32,832 participants in the large-scale cohort study (45 and Up Study), who joined the study from 2006 to 2009 and who were admitted to 313 hospitals in New South Wales, Australia, for at least an overnight stay, up to a year prior to study entry. OUTCOME MEASURES: Agreement between self-report and recording of six morbidities in administrative hospital data, and between-hospital variation and predictors of positive agreement between the two data sources. RESULTS: Agreement between data sources was good for diabetes (κ=0.79); moderate for smoking (κ=0.59); fair for heart disease, stroke and hypertension (κ=0.40, κ=0.30 and κ =0.24, respectively); and poor for obesity (κ=0.09), indicating that a large number of individuals with self-reported morbidities did not have a corresponding diagnosis coded in their hospital records. Significant between-hospital variation was found (ranging from 8% of unexplained variation for diabetes to 22% for heart disease), with higher agreement in public and large hospitals, and hospitals with greater depth of coding. CONCLUSIONS: The recording of six common health conditions in administrative hospital data is highly variable, and for some conditions, very poor. To support more valid performance comparisons, it is important to stratify or control for factors that predict the completeness of recording, including hospital depth of coding and hospital type (public/private), and to increase efforts to standardise recording across hospitals. Studies using these conditions for risk adjustment should also be cautious of their use in smaller hospitals.

Macdonald, K. I., Kilty, S. J. et van Walraven, C. (2016). "Chronic rhinosinusitis identification in administrative databases and health surveys: A systematic review." Laryngoscope **126**(6): 1303-1310.

OBJECTIVES/HYPOTHESIS: Much of the epidemiological data on chronic rhinosinusitis (CRS) are based on large administrative databases and health surveys. The accuracy of CRS identification with these methods is unknown. METHODS: A systematic review was performed to identify studies that measured the accuracy of CRS diagnoses in large administrative databases or within health surveys. The Quality Assessment of Diagnostic Accuracy Studies 2 tool was used to assess study quality. RESULTS: Of 512 abstracts initially identified, 122 were selected for full-text review; only three studies (2.5%) measured the accuracy of CRS patient identification. In a single, large administrative database study with a CRS prevalence of 54.8%, a single International Classification of Diseases-9th Revision diagnostic code for CRS had a positive predictive value (PPV) of only 34%. A diagnostic code algorithm identified CRS patients with a PPV of 91.3% (95% confidence interval [CI], 85.3-95.1); in a population with a CRS prevalence of 5%, this algorithm had a PPV of 31%. In health survey studies having an estimated CRS prevalence of 25% to 46%, self-reported symptom-based CRS diagnosis had a PPV of 62% (95% CI, 50.2-72.1) when nasal endoscopy was the gold standard for CRS diagnosis, and 70% (95% CI, 57.4-80.8) when otolaryngologist-based CRS diagnosis (after interview and nasal endoscopy) was the gold standard. CONCLUSION: Most health administrative data and health surveys examining CRS did not consider the accuracy of case identification. For unselected populations, administrative data and health surveys using self-reported diagnoses inaccurately identify patients with CRS. Epidemiological results based on such data should be interpreted with these results in mind. Laryngoscope, 126:1303-1310, 2016.

Mason, J., Laporte, A., McDonald, J. T., et al. (2023). "Health Reporting from Different Data Sources: Does it Matter for Mental Health?" J Ment Health Policy Econ **26**(1): 33-57.

BACKGROUND: Mental disorders are typically stigmatized conditions associated with negative stereotypes, which may lead individuals to underreport them. Thus, survey data may be subject to biases. Although administrative data has some limitations, it is an alternative data source that may be considered more objective. AIMS OF THE STUDY: This study aimed to identify the degree of agreement between survey and administrative health care data for mental health conditions, factors affecting underreporting, and whether underreporting also occurs for physical health conditions. METHODS: We

used Ontario data from the Canadian Community Health Survey linked to health records to examine the presence of mental health conditions (i.e., schizophrenia and mood disorders) and select physical health conditions (i.e., diabetes and cancer). Using administrative data as the reference standard, we created four categories for each health condition based on the level of agreement between the two data sources: consistent cases and non-cases (i.e. individuals with concordant data based on their reported health condition), and people who were found to underreport and overreport a condition (i.e. where the condition was present in the administrative data, but not in the survey data and vice-versa, respectively). The overall level of agreement was assessed using Cohen's kappa statistic. Probit regressions were estimated to determine the factors affecting underreporting. RESULTS: The Kappa statistics for mood disorder was fair (k= 0.26) and moderate for schizophrenia (k = 0.49). Physical health conditions had higher kappa values (diabetes, k = 0.81; ever having cancer, k = 0.68), with the exception of currently having cancer (k = 0.24). Underreporting was highest for the most stigmatizing condition, schizophrenia (63%), followed by mood disorders (39%) and cancer (39%), and lowest for diabetes (25%). Older age, being born in Africa and Asia, and being employed all increased the probability of underreporting among individuals identified in the administrative data; the opposite held for social assistance. DISCUSSION: We extended previous work on mental health reporting by combining survey data with administrative data to examine the level of agreement between respondents' self-reported mental health and administrative records. The data include some mental disorders not studied previously. We examined the entire adult population; this is important because prevalence of schizophrenia may be less common among older population groups due to higher mortality among this patient population. Additionally, there may be potential age-related differences in stigma and mental health conditions. The administrative health data captured only health services covered by the public provincial health insurance plan and thus did not capture medical care provided by psychologists, social workers, and nurses. While this would affect Kappa statistic values, it does not directly affect the underreporting analyses. IMPLICATIONS FOR HEALTH CARE PROVISION AND USE: Our results suggest that disclosure of mental health conditions may differ by the level of stigma, which has implications for obtaining accurate estimates of mental health prevalence from self-reported data sources.

Mertens, E., Peñalvo, J. et Vandevijvere, S. (2022). "Comparing health insurance and survey data in estimating prevalence of chronic diseases." The European Journal of Public Health **32**(suppl.3): ckac131.145.

Mirel, L. B., Golden, C., Keralis, J. M., et al. (2019). "Evaluating Survey Report of Social Security Disability Benefit Receipt Using Linked National Health Interview Survey and Social Security Administration Data." Natl Health Stat Report(131): 1-15.

Linking nationally representative population health survey data with Social Security Administration (SSA) disability program data provides a rich source of information on program recipients. Survey participant data from the 1998-2005 National Health Interview Survey (NHIS) were linked to SSA administrative records from 1997 through 2005. The goal of this study was to assess agreement between the actual benefit receipt based on the SSA administrative records and the survey report of benefit receipt in the linked NHIS and SSA file for the U.S. civilian noninstitutionalized population. This evaluation provides information on the expected accuracy of survey report of Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) benefit receipt, including how participant characteristics may be associated with reporting misclassification. The results indicate that there is some underreporting of SSA disability benefit receipt based on the NHIS responses compared with the SSA administrative records. The analysis identified some differences between the concordant and discordant groups for selected characteristics, but there were no clear patterns among the different survey questions or the different survey participant characteristics.

Mirel, L. B., Simon, A. E., Golden, C., et al. (2014). "Concordance between survey report of Medicaid enrollment and linked Medicaid administrative records in two national studies." Natl Health Stat Report(72): 1-9.

The National Health and Nutrition Examination Survey (NHANES) and the National Health Interview Survey (NHIS) are population-based surveys that have each been linked to administrative data from the Centers for Medicare and Medicaid Services (CMS): the Medicaid Analytic eXtract (MAX) files.

These linked data were used to examine, among children under age 18 years, respondent-level concordance between Medicaid or the Children's Health Insurance Program (CHIP) enrollment as reported in each survey (NHANES and NHIS) and as indicated by administrative data from the MAX files. Concordance was defined as having Medicaid/CHIP reported as a health insurance source in the survey questionnaire and having a CMS Medicaid/CHIP administrative record in the same month and year as the interview. Records were also considered concordant if there was no report of Medicaid/CHIP coverage based on the interview response and no match to the CMS administrative records for Medicaid enrollment. Between NHANES and MAX, 88% of observations were concordant with respect to Medicaid or CHIP enrollment, yielding a Kappa of 0.71. Between NHIS and MAX, 89% of observations were concordant with respect to Medicaid or CHIP enrollment, yielding a Kappa of 0.73. These concordance rates provide support for the use of both administrative and NHANES and NHIS as a valuable tool for public health researchers and survey methodologists.

Mouly, D., Van Cauteren, D., Vincent, N., et al. (2016). "Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases." Epidemiol Infect **144**(3): 591-601.

Waterborne disease outbreaks (WBDO) of acute gastrointestinal illness (AGI) are a public health concern in France. Their occurrence is probably underestimated due to the lack of a specific surveillance system. The French health insurance database provides an interesting opportunity to improve the detection of these events. A specific algorithm to identify AGI cases from drug payment reimbursement data in the health insurance database has been previously developed. The purpose of our comparative study was to retrospectively assess the ability of the health insurance data to describe WBDO. Data from the health insurance database was compared with the data from cohort studies conducted in two WBDO in 2010 and 2012. The temporal distribution of cases, the day of the peak and the duration of the epidemic, as measured using the health insurance data, were similar to the data from one of the two cohort studies. However, health insurance data accounted for 54 cases compared to the estimated 252 cases accounted for in the cohort study. The accuracy of using health insurance data to describe WBDO depends on the medical consultation rate in the impacted population. As this is never the case, data analysis underestimates the total number of AGI cases. However this data source can be considered for the development of a detection system of a WBDO in France, given its ability to describe an epidemic signal.

Muggah, E., Graves, E., Bennett, C., et al. (2013). "Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report." BMC Public Health **13**: 16.

BACKGROUND: Health administrative data is increasingly being used for chronic disease surveillance. This study explored agreement between administrative and survey data for ascertainment of seven key chronic diseases, using individually linked data from a large population of individuals in Ontario, Canada. METHODS: All adults who completed any one of three cycles of the Canadian Community Health Survey (2001, 2003 or 2005) and agreed to have their responses linked to provincial health administrative data were included. The sample population included 85,549 persons. Previously validated case definitions for myocardial infarction, asthma, diabetes, chronic lung disease, stroke, hypertension and congestive heart failure based on hospital and physician billing codes were used to identify cases in health administrative data and these were compared with self-report of each disease from the survey. Concordance was measured using the Kappa statistic, percent positive and negative agreement and prevalence estimates. RESULTS: Agreement using the Kappa statistic was good or very good (kappa range: 0.66-0.80) for diabetes and hypertension, moderate for myocardial infarction and asthma and poor or fair (kappa range: 0.29-0.36) for stroke, congestive heart failure and COPD. Prevalence was higher in health administrative data for all diseases except stroke and myocardial infarction. Health Utilities Index scores were higher for cases identified by health administrative data compared with self-reported data for some chronic diseases (acute myocardial infarction, stroke, heart failure), suggesting that administrative data may pick up less severe cases. CONCLUSIONS: In the general population, discordance between self-report and administrative data was large for many chronic diseases, particularly disease with low prevalence, and differences were not easily explained by individual and disease characteristics.

Muhajarine, N., Mustard, C., Roos, L. L., et al. (1997). "Comparison of survey and physician claims data for detecting hypertension." J Clin Epidemiol **50**(6): 711-718.

Using linked data from the Manitoba (Canada) Heart Health Survey (MHHS) and physician service claims files we assessed the degree to which self-reported hypertension and clinically measured hypertension agreed with physician claims hypertension, and examined the likely sources of disagreement. The overall agreement between survey and claims data for hypertension detection was moderate to high: 82% (kappa = 0.56) for self-reported and physician claims hypertension, and 85% (kappa = 0.60) for clinically measured and physician claims hypertension. In the comparison between self-report and physician claims, those who were classified as obese, diabetic, or a homemaker were significantly more likely to have a hypertension measure not confirmed by the other. Disagreement between clinically measured and physician claims was also more common among the obese and homemakers, as well as those on medication for heart diseases, elevated cholesterol levels (LDL), and 35 years of age and older. The high overall level of agreement among these three measures suggest that each may be used with confidence as an indication of hypertension; however, the agreement appears lower among individuals presenting a more complicated clinical profile.

Navin Cristina, T. J., Stewart Williams, J. A., Parkinson, L., et al. (2016). "Identification of diabetes, heart disease, hypertension and stroke in mid- and older-aged women: Comparing self-report and administrative hospital data records." Geriatr Gerontol Int **16**(1): 95-102.

AIM: To estimate the prevalence of diabetes, heart disease, hypertension and stroke in self-report and hospital data in two cohorts of women; measure sensitivity and agreement between data sources; and compare between cohorts. METHODS: Women born between 1946-1951 and 1921-1926 who participated in the Australian Longitudinal Study on Women's Health (ALSWH); were New South Wales residents; and admitted to hospital (2004-2008) were included in the present study. The prevalence of diabetes, heart disease, hypertension and stroke was estimated using self-report (case 1 at latest survey, case 2 across multiple surveys) and hospital records. Agreement (kappa) and sensitivity (%) were calculated. Logistic regression measured the association between patient characteristics and agreement. RESULTS: Hypertension had the highest prevalence and estimates were higher for older women: 32.5% case 1, 45.4% case 2, 12.8% in hospital data (1946-1951 cohort); 57.8% case 1, 73.2% case 2, 38.2% in hospital data (1921-1926 cohort). Agreement was substantial for diabetes: kappa = 0.75 case 1, kappa = 0.70 case 2 (1946-1951 cohort); kappa = 0.77 case 1, kappa = 0.80 case 2 (1921-1926 cohort), and lower for other conditions. The 1946-1951 cohort had 2.08 times the odds of agreement for hypertension (95% CI 1.56 to 2.78; P < 0.0001), and 6.25 times the odds of agreement for heart disease (95% CI 4.35 to 10.0; P < 0.0001), compared with the 1921-1926 cohort. CONCLUSION: Substantial agreement was found for diabetes, indicating accuracy of ascertainment using self-report or hospital data. Self-report data appears to be less accurate for heart disease and stroke. Hypertension was underestimated in hospital data. These findings have implications for epidemiological studies relying on self-report or administrative data.

Oberski, D. L., Kirchner, A., Eckman, S., et al. (2017). "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models." Journal of the American Statistical Association **112**(520): 1477-1489.

Administrative data are increasingly important in statistics, but, like other types of data, may contain measurement errors. To prevent such errors from invalidating analyses of scientific interest, it is therefore essential to estimate the extent of measurement errors in administrative data. Currently, however, most approaches to evaluate such errors involve either prohibitively expensive audits or comparison with a survey that is assumed perfect. We introduce the " generalized multitrait-multimethod" GMTMM) model, which can be seen as a general framework for evaluating the quality of administrative and survey data simultaneously. This framework allows both survey and administrative data to contain random and systematic measurement errors. Moreover, it accommodates common features of administrative data such as discreteness, nonlinearity, and nonnormality, improving similar existing models. The use of the GMTMM model is demonstrated by

application to linked survey-administrative data from the German Federal Employment Agency on income from of employment, and a simulation study evaluates the estimates obtained and their robustness to model misspecification. Supplementary materials for this article are available online.

O'Donnell, S., Vanderloo, S., McRae, L., et al. (2016). "Comparison of the estimated prevalence of mood and/or anxiety disorders in Canada between self-report and administrative data." Epidemiol Psychiatr Sci **25**(4): 360-369.

BACKGROUND: To compare trends in the estimated prevalence of mood and/or anxiety disorders identified from two data sources (self-report and administrative). Reviewing, synthesising and interpreting data from these two sources will help identify potential factors that underlie the observed estimates and inform public health action. METHOD: We used self-reported, diagnosed mood and/or anxiety disorder cases from the Canadian Community Health Survey (CCHS) across a 5-year span (from 2003 to 2009) to estimate the prevalence among the Canadian population aged ≥15 years. We also estimated the prevalence of mood and/or anxiety disorders using the Canadian Chronic Disease Surveillance System (CCDSS), which identified cases using ICD-9/-10-CA codes from physician billing claims and hospital discharge records during the same time period. The prevalence rates for mood and/or anxiety disorders were compared across the CCHS and CCDSS by age and sex for all available years of data from 2003 to 2009. Summary rates were age-standardised to the Canadian population as of 1 October 1991. RESULTS: In 2009, the prevalence of mood and/or anxiety disorders was 9.4% using self-reported data v. 11.3% using administrative data. Prevalence rates obtained from administrative data were consistently higher than those from self-report for both men and women. However, due to an increase in the prevalence of self-reported cases, these differences decreased over time (rate ratios for both sexes: 1.6-1.2). Prevalence estimates were consistently higher among females compared with males irrespective of data source. While differences in the prevalence estimates between the two data sources were evident across all age groups, the reduction of these differences was greater among adolescent, young and middle-aged adults compared with those 70 years and older. CONCLUSIONS: The overall narrowing of differences over time reflects a convergence of information regarding the prevalence of mood and/or anxiety disorders trends between self-report and administrative data sources. While the administrative data-based prevalences remained relatively stable, the self-reported prevalences increased over time. These observations may reflect positive societal changes in the perceptions of mental health (declining stigma) and/or increasing mental health literacy. Additional research using non-ecological data is required to further our understanding of the observed findings and trends, including a data linkage exercise permitting a comparison of prevalence estimates and population characteristics from these two data sources both separately and merged.

Olomu, A., Corser, W., Stommel, M., et al. (2012). "Do self-report and medical record comorbidity data predict longitudinal functional capacity and quality of life health outcomes similarly?" BMC Health Serv Res **12**(1): 398-398.

The search for a reliable, valid and cost-effective comorbidity risk adjustment method for outcomes research continues to be a challenge. The most widely used tool, the Charlson Comorbidity Index (CCI) is limited due to frequent missing data in medical records and administrative data. Patient self-report data has the potential to be more complete but has not been widely used. The purpose of this study was to evaluate the performance of the Self-Administered Comorbidity Questionnaire (SCQ) to predict functional capacity, quality of life (QOL) health outcomes compared to CCI medical records data. An SCQ-score was generated from patient interview, and the CCI score was generated by medical record review for 525 patients hospitalized for Acute Coronary Syndrome (ACS) at baseline, three months and eight months post-discharge. Linear regression models assessed the extent to which there were differences in the ability of comorbidity measures to predict functional capacity (Activity Status Index [ASI] scores) and quality of life (EuroQOL 5D [EQ5D] scores). The CCI (R2 = 0.245; p = 0.132) did not predict quality of life scores while the SCQ self-report method (R2 = 0.265; p < 0.0005) predicted the EQ5D scores. However, the CCI was almost as good as the SCQ for predicting the ASI scores at three and six months and performed slightly better in predicting ASI at eight-month follow up (R2 = 0.370; p < 0.0005 vs. R2 = 0.358; p < 0.0005) respectively. Only age, gender, family income and Center for Epidemiologic Studies-Depression (CESD) scores showed significant association with both measures in predicting QOL and functional capacity. Although our model R-squares were fairly low, these results

show that the self-report SCQ index is a good alternative method to predict QOL health outcomes when compared to a CCI medical record score. Both measures predicted physical functioning similarly. This suggests that patient self-reported comorbidity data can be used for predicting physical functional capacity and QOL and can serve as a reliable risk adjustment measure. Self-report comorbidity data may provide a cost-effective alternative method for risk adjustment in clinical research, health policy and organizational improvement analyses. Clinical Trials.gov NCT00416026

Paalanen, L., Koponen, P., Laatikainen, T., et al. (2018). "Public health monitoring of hypertension, diabetes and elevated cholesterol: comparison of different data sources." Eur J Public Health **28**(4): 754-765.

BACKGROUND: Three data sources are generally used in monitoring health on the population level. Health interview surveys (HISs) are based on participants' self-report. Health examination surveys (HESs) yield more objective data, and also persons who are unaware of their elevated risks can be detected. Medical records (MRs) and other administrative registers also provide objective data, but their availability, coverage and quality vary between countries. We summarized studies comparing self-reported data with (i) measured data from HESs or (ii) MRs. We aimed to describe differences in feasibility and comparability of different data sources for monitoring (i) elevated blood pressure or hypertension (ii) elevated blood glucose or diabetes and (iii) elevated total cholesterol. METHODS: We conducted a literature search to identify studies, which validated self-reported measures against objective measures. We found 30 studies published since the year 2000 fulfilling our inclusion criteria (targeted to adults and comparing prevalence among the same persons). RESULTS: Hypertension and elevated total cholesterol were prone to be under-estimated in HISs. The under-estimate was more pronounced, when the HIS data were compared with HES data, and lower when compared with MRs. For diabetes, the HISs and the objective methods resulted in fairly similar prevalence rates. CONCLUSION: The three data sources measure different manifestations of the risk factors and cannot be expected to yield similar prevalence rates. Using HIS data only may lead to under-estimation of elevated risk factor levels or disease prevalence. Whenever possible, information from the three data sources should be evaluated and combined.

Palin, J. L., Goldner, E. M., Koehoorn, M., et al. (2011). "Primary mental health care visits in self-reported data versus provincial administrative records." Health Reports **22**(2): 41-47.

Background Survey data and provincial administrative health data are the major sources of population estimates of mental health care visits to General Practitioners (GPs). Previous research has suggested that self-reported estimates of the number of mental health-related visits per person to health professionals may exceed estimates obtained from physician reimbursement records. Data and methods Self-reported data from the 2002 Canadian Community Health Survey (CCHS): Mental Health and Well-being and administrative records from the Medical Services Plan of British Columbia were linked. The analytic sample consisted of 145 CCHS respondents who had at least one mental health visit to a GP in the past 12 months according to both data sources. High Reporters (self-reported visits exceeded number in administrative data), Low Reporters (self-reported visits were less than number in administrative data), and Exact Matches were analyzed in two ways. The first analysis used diagnostic codes to identify mental health-related visits in the administrative data. For the second analysis, all GP visits in the administrative data were counted as "possibly" mental health-related. Differences were described based on the median number of visits. Results When diagnostic codes were used to identify mental-health-related visits in the administrative data, High Reporters (49%) substantially exceeded Low Reporters (24%). The remaining 27% were Exact Matches. Based on a broader definition of a mental health visit, 51% were Exact Matches. High reporting was common among people with mental disorders. Interpretation Self-reported data and administrative data provide different estimates of the number of mental health visits per person to GPs. The discrepancy can be large.

Palmer, L., Johnston, S. S., Rousculp, M. D., et al. (2012). "Agreement between Internet-based self- and proxy-reported health care resource utilization and administrative health care claims." Value Health **15**(3): 458-465.

OBJECTIVES: Although Internet-based surveys are becoming more common, little is known about agreement between administrative claims data and Internet-based survey self- and proxy-reported

health care resource utilization (HCRU) data. This analysis evaluated the level of agreement between self- and proxy-reported HCRU data, as recorded through an Internet-based survey, and administrative claims-based HCRU data. METHODS: The Child and Household Influenza-Illness and Employee Function study collected self- and proxy-reported HCRU data monthly between November 2007 and May 2008. Data included the occurrence and number of visits to hospitals, emergency departments, urgent care centers, and outpatient offices for a respondent's and his or her household members' care. Administrative claims data from the MarketScan® Databases were assessed during the same time and evaluated relative to survey-based metrics. Only data for individuals with employer-sponsored health care coverage linkable to claims were included. The Kappa (κ) statistic was used to evaluate visit concordance, and the intraclass correlation coefficient was used to describe frequency consistency. RESULTS: Agreement for presence of a health care visit and the number of visits were similar for self- and proxy-reported HCRU data. There was moderate to substantial agreement related to health care visit occurrence between survey-based and claims-based HCRU data for inpatient, emergency department, and office visits (κ: 0.47-0.77). There was less agreement on health care visit frequencies, with intraclass correlation coefficient values ranging from 0.14 to 0.71. CONCLUSIONS: This study's agreement values suggest that Internet-based surveys are an effective method to collect self- and proxy-reported HCRU data. These results should increase confidence in the use of the Internet for evaluating disease burden.

Parkinson, L., Curryer, C., Gibberd, A., et al. (2013). "Good agreement between self-report and centralized hospitalizations data for arthritis-related surgeries." J Clin Epidemiol **66**(10): 1128-1134.

Objectives: To examine the level of agreement between self-reported and hospital administration records of arthritis-related surgeries for two large samples of community-dwelling older women in Australia, born between 1921-1926 and 1946-1951. Study Design and Setting: Self-report survey data from the Australian Longitudinal Study on Women's Health was linked to inpatient hospital data from the New South Wales Admitted Patient Data Collection. Levels of agreement were compared using Cohen's kappa, sensitivity, specificity, and positive and negative predictive values. Reasons for false positives were examined. Results: This study found good agreement (kappa >0.70; sensitivity 'and specificity >0.80) between self-report and hospitalizations data for arthritis-related surgeries. Conclusions: This study provides new evidence for good agreement between self-reported health survey data and administrative records of arthritis-related joint procedures, and supports the use of self-report surveys in epidemiological studies of joint procedures where administrative data are either not available or not readily accessible, or where more extensive contextual information is needed. The use of health survey data in conjunction with administrative data has an important role to play in public health planning and policy. (C) 2013 Elsevier Inc. All rights reserved.

Payette, Y., de Moura, C. S., Boileau, C., et al. (2020). "Is there an agreement between self-reported medical diagnosis in the CARTaGENE cohort and the Québec administrative health databases?" Int J Popul Data Sci **5**(1): 1155.

BACKGROUND: Population health studies often use existing databases that are not necessarily constituted for research purposes. The question arises as to whether different data sources such as in administrative health data (AHD) and self-report questionnaires are equivalent and lead to similar information. OBJECTIVES: The main objective of this study was to assess the level of agreement between self-reported medical conditions and medical diagnosis captured in AHD. A secondary objective was to identify predictors of agreement among medical conditions between the two data sources. Therefore, the purposes of the study were to explore the extent to which these two methods of commonly used public health data collection provide concordant records and identify the main predictors of statistical variations. METHODS: Data were extracted from CARTaGENE, a population-based cohort in Québec, Canada, which was linked to the provincial health insurance records of the same individuals, namely the MED-ÉCHO database from the Régie de l'assurance maladie du Québec (RAMQ) and the fee-for-service billing records provided by the physician, for the time period 1998-2012. Agreement statistics (kappa coefficient) along with sensitivity, specificity and predictive positive value were calculated for 19 chronic conditions and 12 types of cancers. Logistic regressions were used to identify predictors of concordance between self-report and AHD from significant covariates (sex,

age groups, education, region, income, heavy utilization of health care system and Charlson comorbidity index). RESULTS: Agreement between self-reported data and AHD across diseases ranged from kappa of 0.09 for chronic renal failure to 0.86 for type 2 diabetes. Sensitivity of self-reported data was higher than 50% for 14 out of the 31 medical conditions studied, especially for myocardial infarction (88.62%), breast cancer (86.28%), and diabetes (85.06%). Specificity was generally high with a minimum value of 89.70%. Lower concordance between data sources was observed for higher frequency of health care utilization and higher comorbidity scores. CONCLUSION: Overall, there was moderate agreement between the two data sources but important variations were found depending on the type of disease. This suggests that CARTaGENE's participants were generally able to correctly identify the kind of diseases they suffer from, with some exceptions. These results may help researchers choose adequate data sources according to specific study objectives. These results also suggest that Québec's AHD seem to underestimate the prevalence of some chronic conditions, which might result in inaccurate estimates of morbidity with consequences for public health surveillance.

Plante, C., Goudreau, S., Jacques, L., et al. (2014). "Agreement between survey data and Régie de l'assurance maladie du Québec (RAMQ) data with respect to the diagnosis of asthma and medical services use for asthma in children." Chronic Dis Inj Can **34**(4): 256-262.

INTRODUCTION: The goal of this study was to assess the agreement between the results of a respiratory health survey conducted in Montréal on children aged 6 months to 12 years and the Régie de l'assurance maladie du Québec (RAMQ, Quebec health insurance board) database in terms of the diagnosis of asthma and medical services use. A secondary aim was to evaluate the effect of the survey method used (Internet-based survey or telephone survey). METHODS: We assessed whether a diagnosis of asthma was made for 7922 children. In addition, we compared the use of medical services for asthma (emergency department visits and hospitalizations) in the 12 months preceding the survey for the 402 children considered to have asthma, using 2 groups of respiratory diagnoses and 2 data linkage periods. The agreement between the 2 data sources was evaluated using the kappa statistic (κ) and sensitivity and specificity, as well as percentages of agreement, overreporting and under-reporting with respect to health services use. RESULTS: Moderate agreement was found between the 2 data sources (survey and RAMQ data) in terms of the diagnosis of asthma (κ = 0.54 and κ = 0.60 depending on the definition used). Specificity was high (93% and 96%), but sensitivity varied (50% and 65%). Respondents over-reported health services use, resulting in moderate kappa values (0.49 for emergency department visits and 0.48 for hospitalizations). However, when more diagnoses were included in the definition and when the linkage period was extended (15 rather than 12 months), the kappa values increased (0.59 for emergency department visits and 0.64 for hospitalizations) and sensitivity and specificity were high. Slightly higher agreement was obtained for the Internet-based survey relative to the telephone survey. CONCLUSION: The findings validate the use of survey data with respect to the diagnosis of pediatric asthma and major health services use for this disease.

Podmore, B., Hutchings, A., Konan, S., et al. (2019). "The agreement between chronic diseases reported by patients and derived from administrative data in patients undergoing joint arthroplasty." BMC Med Res Methodol **19**(1): 87.

BACKGROUND: This study examined the agreement between patient-reported chronic diseases and hospital administrative records in hip or knee arthroplasty patients in England. METHODS: Survey data reported by 676,428 patients for the English Patient Reported Outcome Measures (PROMs) programme was linked to hospital administrative data. Sensitivity and specificity of 11 patient-reported chronic diseases were estimated with hospital administrative data as reference standard. RESULTS: Specificity was high (> 90%) for all 11 chronic diseases. However, sensitivity varied by disease with the highest found for 'diabetes' (87.5%) and 'high blood pressure' (74.3%) and lowest for 'kidney disease' (18.8%) and 'leg pain due to poor circulation' (26.1%). Sensitivity was increased for diseases that were given as specific examples in the questionnaire (e.g. 'parkinson's disease' (65.6%) and 'multiple sclerosis' (69.5%), compared to 'diseases of the nervous system' (20.9%)). CONCLUSIONS: Patients can give information about the presence of chronic diseases that is consistent with chronic diseases derived from hospital administrative data if the description in the patient questionnaire is precise and if the disease is familiar to most patients and has significant impact on their life. Such

patient questionnaires need to be validated before they are used for research and service evaluation projects.

Raghunathan, T., Ghosh, K., Rosen, A., et al. (2020). "Combining Information from Multiple Data Sources to Assess Population Health." J Surv Stat Methodol **9**(3): 598-625.

Information about an extensive set of health conditions on a well-defined sample of subjects is essential for assessing population health, gauging the impact of various policies, modeling costs, and studying health disparities. Unfortunately, there is no single data source that provides accurate information about health conditions. We combine information from several administrative and survey data sets to obtain model-based dummy variables for 107 health conditions (diseases, preventive measures, and screening for diseases) for elderly (age 65 and older) subjects in the Medicare Current Beneficiary Survey (MCBS) over the fourteen-year period, 1999-2012. The MCBS has prevalence of diseases assessed based on Medicare claims and provides detailed information on all health conditions but is prone to underestimation bias. The National Health and Nutrition Examination Survey (NHANES), on the other hand, collects self-reports and physical/laboratory measures only for a subset of the 107 health conditions. Neither source provides complete information, but we use them together to derive model-based corrected dummy variables in MCBS for the full range of existing health conditions using a missing data and measurement error model framework. We create multiply imputed dummy variables and use them to construct the prevalence rate and trend estimates. The broader goal, however, is to use these corrected or modeled dummy variables for a multitude of policy analysis, cost modeling, and analysis of other relationships either using them as predictors or as outcome variables.

Raina, P., Torrance-Rynard, V., Wong, M., et al. (2002). "Agreement between self-reported and routinely collected health-care utilization data among seniors." Health Serv Res **37**(3): 751-774.

OBJECTIVE: To examine the agreement between self-reported and routinely collected administrative health-care utilization data, and the factors associated with agreement between these two data sources. DATA SOURCES/STUDY SETTING: A representative sample of seniors living in an Ontario county within Canada was identified using the Ontario Ministry of Health's Registered Persons Data Base in 1992. Health professional billing information and hospitalization data were obtained from the Ontario Ministry of Health and Long-Term Care (OMH) and the Ontario Health Insurance Plan (OHIP). STUDY DESIGN: A cross-sectional survey was carried out to assess any contact and frequency of contacts with health professionals and hospital admissions. Similar information was obtained from routinely collected administrative data. The level of agreement was assessed using the proportion of absolute agreement, Cohen's kappa statistic (kappa), and the intraclass correlation coefficient (ICC). Logistic and linear regressions were used to identify factors that were associated with the magnitude and direction of disagreement respectively. DATA COLLECTION/EXTRACTION METHODS: Telephone interviews were conducted on 1,054 seniors, and complete data were available for 1,038 seniors. Each respondent's personal health number was used to electronically link survey data with health professional billing and hospitalization databases. PRINCIPAL FINDINGS: Substantial to almost perfect agreement was found for the contact utilization measures, while agreement on volume utilization measures varied from poor to almost perfect. In surveys, seniors overreported contact with general practitioners and physiotherapists or chiropractors, and underreported contact with other medical specialists. Seniors also underreported the number of contacts with general practitioners and other medical specialists. The odds of agreement decreased if respondents were male, aged 75 years and older, had incomes of less than $25,000, had poor/fair/good self-assessed health status, or had two or more chronic conditions. CONCLUSION: The findings of this study indicate that there are substantial discrepancies between self-reported and administrative data among older adults. Researchers seeking to examine health-care use among older adults need to consider these discrepancies in the interpretation of their results. Failure to recognize these discrepancies between survey and administrative data among older adults may lead to the establishment of inappropriate health-care policies.

Rammon, J., He, Y. et Parker, J. D. (2018). "Accounting for study participants who are ineligible for linkage: a multiple imputation approach to analyzing the linked National Health and Nutrition Examination Survey and Centers for Medicare and Medicaid Services' Medicaid data." Health Serv Outcomes Res Methodol **19**(2-3): 87-105.

Data from the National Health and Nutrition Examination Survey have been linked to the Center for Medicare and Medicaid Services' Medicaid Enrollment and Claims Files for the survey years 1999-2004. The linked data are produced by the National Center for Health Statistics' (NCHS) Data Linkage Program and are available in the NCHS Research Data Center. This project compares the usefulness of multiple imputation to account for data linkage ineligibility and other survey nonresponse with currently recommended weight adjustment procedures. Estimated differences in environmental smoke exposure across Medicaid/Children's Health Insurance Program (CHIP) enrollment status among children ages 3-15 years are examined as a motivating example. Comparisons are drawn across the three different estimates: one that uses MI to impute the administrative Medicaid/CHIP status of those who are ineligible for linkage, a second that uses the linked data restricted to linkage eligible participants with a basic weight adjustment, and a third that uses self-reported Medicaid/CHIP status from the survey data. The results indicate that estimates from the multiple imputation analysis were comparable to those found when using weight adjustment procedures and had the added benefit of incorporating all survey participants (linkage eligible and linkage ineligible) into the analysis. We conclude that both multiple imputation and weight adjustment procedures can effectively account for survey participants who are ineligible for linkage.

Robinson, J. R., Young, T. K., Roos, L. L., et al. (1997). "Estimating the burden of disease. Comparing administrative data and self-reports." Med Care **35**(9): 932-947.

OBJECTIVES: A cardiovascular health survey of a representative sample of the adult population of Manitoba, Canada was combined with the provincial health insurance claims database to determine the accuracy of survey questions in detecting cases of diabetes, hypertension, ischemic heart disease, stroke, and hypercholesterolemia. METHODS: Of 2,792 subjects in the survey, 97.7% were linked successfully using a scrambled personal health insurance number. Hospital and physician claims were extracted for these individuals for the 3-year period before the survey. RESULTS: The authors found no benefits to using restrictive criteria for entrance into the study (ie, requiring more than one diagnosis to define a case). Using additional years of data increased agreement between data sources. Kappa values indicated high levels of agreement between administrative data and self-reports for diabetes (0.72) and hypertension (0.59); kappa values were approximately 0.4 for the other conditions. Using administrative data as the "gold standard," specificity was generally very high, although cases with hypertension and hypercholesterolemia (diagnosed primarily by laboratory or physical measurement) were associated with a lower specificity than the other conditions. Sensitivity varied markedly and was lowest for "other heart disease" and "stroke". For diabetes and hypertension, inclusion criteria calling for more than one diagnosis reduced the accuracy of case identification, whereas increasing the number of years of data increased accuracy of identification. For diabetes and hypertension, self-reports were fairly accurate in detecting "true" past history of the illness based on physician diagnosis recorded on insurance claims. CONCLUSIONS: This study demonstrates the feasibility of linking a large health survey with administrative data and the validity of self-reports in estimating the prevalence of chronic diseases, especially diabetes and hypertension. A linked data set offers unusual opportunities for epidemiologic and health services research in a defined population.

Rousseau, M. C., Conus, F., El-Zein, M., et al. (2023). "Ascertaining asthma status in epidemiologic studies: a comparison between administrative health data and self-report." BMC Med Res Methodol **23**(1): 201.

BACKGROUND: Studies have suggested that agreement between administrative health data and self-report for asthma status ranges from fair to good, but few studies benefited from administrative health data over a long period. We aimed to (1) evaluate agreement between asthma status ascertained in administrative health data covering a period of 30 years and from self-report, and (2) identify determinants of agreement between the two sources. METHODS: We used administrative health data (1983-2012) from the Quebec Birth Cohort on Immunity and Health, which included

81,496 individuals born in the province of Quebec, Canada, in 1974. Additional information, including self-reported asthma, was collected by telephone interview with 1643 participants in 2012. By design, half of them had childhood asthma based on health services utilization. Results were weighted according to the inverse of the sampling probabilities. Five algorithms were applied to administrative health data (having ≥ 2 physician claims over a 1-, 2-, 3-, 5-, or 30-year interval or ≥ 1 hospitalization), to enable comparisons with previous studies. We estimated the proportion of overall agreement and Kappa, between asthma status derived from algorithms and self-reports. We used logistic regression to identify factors associated with agreement. RESULTS: Applying the five algorithms, the prevalence of asthma ranged from 49 to 55% among the 1643 participants. At interview (mean age = 37 years), 49% and 47% of participants respectively reported ever having asthma and asthma diagnosed by a physician. Proportions of agreement between administrative health data and self-report ranged from 88 to 91%, with Kappas ranging from 0.57 (95% CI: 0.52-0.63) to 0.67 (95% CI: 0.62-0.72); the highest values were obtained with the [≥ 2 physician claims over a 30-year interval or ≥ 1 hospitalization] algorithm. Having sought health services for allergic diseases other than asthma was related to lower agreement (Odds ratio = 0.41; 95% CI: 0.25-0.65 comparing ≥ 1 health services to none). CONCLUSIONS: These findings indicate good agreement between asthma status defined from administrative health data and self-report. Agreement was higher than previously observed, which may be due to the 30-year lookback window in administrative data. Our findings support using both administrative health data and self-report in population-based epidemiological studies.

Sakshaug, J. W., Weir, D. R. et Nicholas, L. H. (2014). "Identifying diabetics in Medicare claims and survey data: implications for health services research." BMC Health Serv Res **14**.

Background: Diabetes health services research often utilizes secondary data sources, including survey self-report and Medicare claims, to identify and study the diabetic population, but disagreement exists between these two data sources. We assessed agreement between the Chronic Condition Warehouse diabetes algorithm for Medicare claims and self-report measures of diabetes. Differences in healthcare utilization outcomes under each diabetes definition were also explored. Methods: Claims data from the Medicare Beneficiary Annual Summary File were linked to survey and blood data collected from the 2006 Health and Retirement Study. A Hemoglobin A1c reading, collected on 2,028 respondents, was used to reconcile discrepancies between the self-report and Medicare claims measures of diabetes. T-tests were used to assess differences in healthcare utilization outcomes for each diabetes measure. Results: The Chronic Condition Warehouse (CCW) algorithm yielded a higher rate of diabetes than respondent self-reports (27.3 vs. 21.2, $p < 0.05$). A1c levels of discordant claims-based diabetics suggest that these patients are not diabetic, however, they have high rates of healthcare spending and utilization similar to diabetics. Conclusions: Concordance between A1c and self-reports was higher than for A1c and the CCW algorithm. Accuracy of self-reports was superior to the CCW algorithm. False positives in the claims data have similar utilization profiles to diabetics, suggesting minimal bias in some types of claims-based analyses, though researchers should consider sensitivity analysis across definitions for health services research.

Short, M. E., Goetzel, R. Z., Pei, X., et al. (2009). "How accurate are self-reports? Analysis of self-reported health care utilization and absence when compared with administrative data." J Occup Environ Med **51**(7): 786-796.

OBJECTIVE: To determine the accuracy of self-reported health care utilization and absence reported on health risk assessments against administrative claims and human resource records. METHODS: Self-reported values of health care utilization and absenteeism were analyzed for concordance to administrative claims values. Percent agreement, Pearson's correlations, and multivariate logistic regression models examined the level of agreement and characteristics of participants with concordance. RESULTS: Self-report and administrative data showed greater concordance for monthly compared with yearly health care utilization metrics. Percent agreement ranged from 30% to 99% with annual doctor visits having the lowest percent agreement. Younger people, males, those with higher education, and healthier individuals more accurately reported their health care utilization and absenteeism. CONCLUSIONS: Self-reported health care utilization and absenteeism may be used as a proxy when medical claims and administrative data are unavailable, particularly for shorter recall periods.

Southern, D. A., Rouleau, C., Wilton, S. B., et al. (2024). "Assessing agreement between population-level administrative pharmaceutical databases and patient-reported medication dispensation in cardiac rehabilitation patients." J Epidemiol Popul Health **72**(5): 202764.

BACKGROUND: Pharmacoepidemiology has emerged as a crucial field in evaluating the use and effects of medications in large populations to ensure their safe and effective use. This study aimed to assess the agreement of cardiac medication use between a provincial medication database, the Pharmaceutical Information Network (PIN), and reconciled medication data from confirmation through patient interviews for patients referred to cardiac rehabilitation. METHODS: The study included data from patients referred to the TotalCardiology Rehabilitation CR program, and medication data was available in both TotalCardiology Rehabilitation charts and PIN. The accuracy of medication data obtained from patient interviews was compared to that obtained from PIN with proportions and kappa statistics to evaluate the reliability of PIN data in assessing medication use. RESULTS: Patient-reported usage was higher for statins (41.6 %) vs. 38.4 %), ACE/ARB, beta-blockers (75.7 %) vs. 73.7 %), DOAC (3.5 %) vs. 2.6 %), and ADP-receptor antagonists (71.0 %) vs. 68.1 %) than if PIN was used. Patient-reported usage data was lower for Ezetimibe (4.7 vs. 4.8 %), Aldosterone antagonists (5.4 %) vs. 5.5 %), digoxin (0.9 %) vs. 1.0 %), calcium channel blockers (19.2 vs. 19.9 %) and warfarin (7.2 %) vs. 8.1 %). The results indicated that the differences between the two sources were very small, with an average agreement of 95.3 % and a kappa of 0.70. CONCLUSION: The study's results, which show a high level of agreement between PIN and patient self-reporting, affirm the reliability of PIN data as a source for obtaining an accurate assessment of medication use. This finding is crucial in the context of pharmacoepidemiology research, where the accuracy of data is paramount. Further research to explore the complementary use of both data sources will be valuable.

St Clair, P., Gaudette, E., Zhao, H., et al. (2017). "Using Self-reports or Claims to Assess Disease Prevalence: It's Complicated." Med Care **55**(8): 782-788.

BACKGROUND: Two common ways of measuring disease prevalence include: (1) using self-reported disease diagnosis from survey responses; and (2) using disease-specific diagnosis codes found in administrative data. Because they do not suffer from self-report biases, claims are often assumed to be more objective. However, it is not clear that claims always produce better prevalence estimates. OBJECTIVE: Conduct an assessment of discrepancies between self-report and claims-based measures for 2 diseases in the US elderly to investigate definition, selection, and measurement error issues which may help explain divergence between claims and self-report estimates of prevalence. DATA: Self-reported data from 3 sources are included: the Health and Retirement Study, the Medicare Current Beneficiary Survey, and the National Health and Nutrition Examination Survey. Claims-based disease measurements are provided from Medicare claims linked to Health and Retirement Study and Medicare Current Beneficiary Survey participants, comprehensive claims data from a 20% random sample of Medicare enrollees, and private health insurance claims from Humana Inc. METHODS: Prevalence of diagnosed disease in the US elderly are computed and compared across sources. Two medical conditions are considered: diabetes and heart attack. RESULTS: Comparisons of diagnosed diabetes and heart attack prevalence show similar trends by source, but claims differ from self-reports with regard to levels. Selection into insurance plans, disease definitions, and the reference period used by algorithms are identified as sources contributing to differences. CONCLUSIONS: Claims and self-reports both have strengths and weaknesses, which researchers need to consider when interpreting estimates of prevalence from these 2 sources.

Svedberg, P., Ropponen, A., Lichtenstein, P., et al. (2010). "Are self-report of disability pension and long-term sickness absence accurate? Comparisons of self-reported interview data with national register data in a Swedish twin cohort." BMC Public Health **10**.

Background: Self-reported disability pension (DP) and sickness absence are commonly used in epidemiological and other studies as a measure of exposure or even as an outcome. The aims were (1) to compare such self-reports with national register information in order to evaluate the validity of self-reported DP and sickness absence, and (2) to estimate the concordance of reporting behaviour in

different twin zygosity groups, also by sex. Methods: All Swedish twins born 1933-1958 who participated in the Screening Across the Lifespan Twin study (SALT) 1998-2003, were included (31,122 individuals). The self-reported DP and long-term sickness absence (LTSA) at the time of interview was compared to the corresponding register information retrieved from the National Social Insurance Agency by calculating the proportions of agreements, kappa, sensitivity, specificity, concordance rates, and chi-square test, to evaluate construct validity. Results: The proportions of overall agreement were 96% and specificity 99% for both DP and LTSA, while the sensitivity was 70% for DP and 45% for LTSA. Kappa estimates were 0.76 for DP, and 0.58 for LTSA. The proportions of positive agreement were 64% for DP and 42% for LTSA. No difference in response style was found between zygosity groups among complete twin pairs for DP and LTSA. Results were similar for women and men and across age. Kappa estimates for DP differed somewhat depending on years of education, 0.68 (college/university) vs. 0.77 (less than 13 years in school) but not for LTSA. Conclusions: Self-reported DP data may be very useful in studies when register information is not available, however, register data is preferred especially for LTSA. The same degree of twin similarity was found for truthful self-report of DP and LTSA in both monozygotic and dizygotic twin pairs. Thus, the response style was not influenced by genetic factors. One consequence of this would be that when estimating the relative importance of genetic and environmental effects from twin models, heritability estimates would not be biased.

Tolonen, H., Reinikainen, J., Koponen, P., et al. (2021). "Cross-national comparisons of health indicators require standardized definitions and common data sources." Arch Public Health **79**(1): 208.

BACKGROUND: Health indicators are used to monitor the health status and determinants of health of the population and population sub-groups, identify existing or emerging health problems which would require prevention and health promotion activities, help to target health care resources in the most adequate way as well as for evaluation of the success of public health actions both at the national and international level. The quality and validity of the health indicator depends both on available data and used indicator definition. In this study we will evaluate existing knowledge about comparability of different data sources for definition of health indicators, compare how selected health indicators presented in different international databases possibly differ, and finally, present the results from a case study from Finland on comparability of health indicators derived from different data sources at national level. METHODS: For comparisons, four health indicators were selected that were commonly available in international databases and available for the Finnish case study. These were prevalence of obesity, hypertension, diabetes, and asthma in the adult populations. Our evaluation has three parts: 1) a scoping review of the latest literature, 2) comparison of the prevalences presented in different international databases, and 3) a case study using data from Finland. RESULTS: Literature shows that comparability of estimated outcomes for health indicators using different data sources such as self-reported questionnaire data from surveys, measured data from surveys or data from administrative health registers, varies between indicators. Also, the case study from Finland showed that diseases which require regular health care visits such as diabetes, comparability is high while for health outcomes which can remain asymptomatic for a long time such as hypertension, comparability is lower. In different international health related databases, country specific results differ due to variations in the used data sources but also due to differences in indicator definitions. CONCLUSIONS: Reliable comparison of the health indicators over time and between regions within a country or across the countries requires common indicator definitions, similar data sources and standardized data collection methods.

Vasquez, M. S., Mertens, E., Berete, F., et al. (2023). "Comparing self-reported health interview survey and pharmacy billing data in determining the prevalence of diabetes, hypertension, and hypercholesterolemia in Belgium." Arch Public Health **81**(1): 121.

BACKGROUND: Administrative and health surveys are used in monitoring key health indicators in a population. This study investigated the agreement between self-reported disease status from the Belgian Health Interview Survey (BHIS) and pharmaceutical insurance claims extracted from the Belgian Compulsory Health Insurance (BCHI) in ascertaining the prevalence of diabetes, hypertension, and hypercholesterolemia. METHODS: Linkage was made between the BHIS 2018 and the BCHI 2018, from which chronic condition was ascertained using the Anatomical Therapeutic Chemical (ATC)

classification and defined daily dose. The data sources were compared using estimates of disease prevalence and various measures of agreement and validity. Multivariable logistic regression was performed for each chronic condition to identify the factors associated to the agreement between the two data sources. RESULTS: The prevalence estimates computed from the BCHI and the self-reported disease definition in BHIS, respectively, are 5.8% and 5.9% diabetes cases, 24.6% and 17.6% hypertension cases, and 16.2% and 18.1% of hypercholesterolemia cases. The overall agreement and kappa coefficient between the BCHI and the self-reported disease status is highest for diabetes and is equivalent to 97.6% and 0.80, respectively. The disagreement between the two data sources in ascertaining diabetes is associated with multimorbidity and older age categories. CONCLUSION: This study demonstrated the capability of pharmacy billing data in ascertaining and monitoring diabetes in the Belgian population. More studies are needed to assess the applicability of pharmacy claims in ascertaining other chronic conditions and to evaluate the performance of other administrative data such as hospital records containing diagnostic codes.

Vinko, M., Kukec, A. et Zaletel-Kragelj, L. (2025). "Mind the Gap: A Retrospective Study of Discrepancies in Self-Reported and Administrative Database-Identified Mental Health Issues in Slovenia." Zdr Varst **64**(3): 143-151.

BACKGROUND: This study assessed discrepancies between self-reported and administrative data sources in identifying mental health issues in Slovenia, and investigated associated socio-demographic factors. METHODS: Data were linked from the 2019 Slovenian European Health Interview Survey (EHIS; n=9,900) and national health administrative databases capturing inpatient hospitalisations, outpatient prescription drugs and mental health-related sick leave. Mental health issues were identified in EHIS by self-report and in administrative databases using diagnostic codes and medication claims. Socio-demographic factors were obtained from EHIS. Discrepancies were assessed and multinomial logistic regression was used to analyse the association between these factors and the source of case identification. RESULTS: Of the 9,900 EHIS respondents, 1,336 (13.5%) self-reported mental health issues, while 1,675 (16.9%) were identified in administrative databases. Only 613 individuals (4.6% of the total sample) were identified in both sources. Older age was associated with being identified in both data sources and administrative data only compared to not being identified. Females and unemployed persons were more likely than males and employed persons to be identified as having mental health issues, regardless of the data source. Compared to those with primary education or lower, individuals with higher education were less likely to be identified in administrative data only or in both data sources. CONCLUSIONS: discrepancies exist between self-reported and administrative data sources in identifying mental health issues. Discrepancies are associated with socio-demographic factors and may lead to different interpretations of population mental health. This study underscores the importance of cautiously interpreting self-reported and administrative health data in public health.

Violán, C., Foguet-Boreu, Q., Hermosilla-Pérez, E., et al. (2013). "Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity." BMC Public Health **13**(1): 251-251.

Health surveys (HS) are a well-established methodology for measuring the health status of a population. The relative merit of using information based on HS versus electronic health records (EHR) to measure multimorbidity has not been established. Our study had two objectives: 1) to measure and compare the prevalence and distribution of multimorbidity in HS and EHR data, and 2) to test specific hypotheses about potential differences between HS and EHR reporting of diseases with a symptoms-based diagnosis and those requiring diagnostic testing. Cross-sectional study using data from a periodic HS conducted by the Catalan government and from EHR covering 80% of the Catalan population aged 15 years and older. We determined the prevalence of 27 selected health conditions in both data sources, calculated the prevalence and distribution of multimorbidity (defined as the presence of ≥2 of the selected conditions), and determined multimorbidity patterns. We tested two hypotheses: a) health conditions requiring diagnostic tests for their diagnosis and management would be more prevalent in the EHR; and b) symptoms-based health problems would be more prevalent in the HS data. We analysed 15,926 HS interviews and 1,597,258 EHRs. The profile of the EHR sample was 52% women, average age 47 years (standard deviation: 18.8), and 68% having at least one of the selected health conditions, the 3 most prevalent being hypertension (20%), depression or anxiety

(16%) and mental disorders (15%). Multimorbidity was higher in HS than in EHR data (60% vs. 43%, respectively, for ages 15-75+, P <0.001, and 91% vs. 83% in participants aged ≥65 years, P <0.001). The most prevalent multimorbidity cluster was cardiovascular. Circulation disorders (other than varicose veins), chronic allergies, neck pain, haemorrhoids, migraine or frequent headaches and chronic constipation were more prevalent in the HS. Most symptomatic conditions (71%) had a higher prevalence in the HS, while less than a third of conditions requiring diagnostic tests were more prevalent in EHR. Prevalence of multimorbidity varies depending on age and the source of information. The prevalence of self-reported multimorbidity was significantly higher in HS data among younger patients; prevalence was similar in both data sources for elderly patients. Self-report appears to be more sensitive to identifying symptoms-based conditions. A comprehensive approach to the study of multimorbidity should take into account the patient perspective.

Olomu, A., Corser, W., Stommel, M., et al. (2012). "Do self-report and medical record comorbidity data predict longitudinal functional capacity and quality of life health outcomes similarly?" BMC Health Serv Res **12**(1): 398-398.

The search for a reliable, valid and cost-effective comorbidity risk adjustment method for outcomes research continues to be a challenge. The most widely used tool, the Charlson Comorbidity Index (CCI) is limited due to frequent missing data in medical records and administrative data. Patient self-report data has the potential to be more complete but has not been widely used. The purpose of this study was to evaluate the performance of the Self-Administered Comorbidity Questionnaire (SCQ) to predict functional capacity, quality of life (QOL) health outcomes compared to CCI medical records data. An SCQ-score was generated from patient interview, and the CCI score was generated by medical record review for 525 patients hospitalized for Acute Coronary Syndrome (ACS) at baseline, three months and eight months post-discharge. Linear regression models assessed the extent to which there were differences in the ability of comorbidity measures to predict functional capacity (Activity Status Index [ASI] scores) and quality of life (EuroQOL 5D [EQ5D] scores). The CCI (R2 = 0.245; p = 0.132) did not predict quality of life scores while the SCQ self-report method (R2 = 0.265; p < 0.0005) predicted the EQ5D scores. However, the CCI was almost as good as the SCQ for predicting the ASI scores at three and six months and performed slightly better in predicting ASI at eight-month follow up (R2 = 0.370; p < 0.0005 vs. R2 = 0.358; p < 0.0005) respectively. Only age, gender, family income and Center for Epidemiologic Studies-Depression (CESD) scores showed significant association with both measures in predicting QOL and functional capacity. Although our model R-squares were fairly low, these results show that the self-report SCQ index is a good alternative method to predict QOL health outcomes when compared to a CCI medical record score. Both measures predicted physical functioning similarly. This suggests that patient self-reported comorbidity data can be used for predicting physical functional capacity and QOL and can serve as a reliable risk adjustment measure. Self-report comorbidity data may provide a cost-effective alternative method for risk adjustment in clinical research, health policy and organizational improvement analyses. Clinical Trials.gov NCT00416026

Wong, J. J., Côté, P., Tricco, A. C., et al. (2021). "Assessing the validity of health administrative data compared to population health survey data for the measurement of low back pain." Pain **162**(1): 219-226.

Low back pain (LBP) is a high-burden condition that lacks routine surveillance data. Health administrative data may be used for surveillance, but their validity for measuring LBP in the general population has not been established. We aimed to (1) determine the validity of health administrative data to measure LBP compared to self-reported LBP in a population-based sample of Ontario adults; and (2) describe the differences in characteristics of LBP cases based on data sources. Adult respondents (≥18 years) of the Canadian Community Health Survey (CCHS) from 2003 to 2012 were included (N = 150,695). Canadian Community Health Survey data were individually linked to health administrative data, including Ontario Health Insurance Plan and hospitalization data. The reference standard was defined as self-reported back problem diagnosed by a health professional in the CCHS. Measurement of LBP from billing records was defined as ≥1 physician billing or procedural code for LBP during the year preceding CCHS interview date. We measured concurrent validity by comparing prevalence, agreement (kappa), and accuracy (sensitivity, specificity, and positive and negative predictive values [PV]) of administrative data to measure LBP. Prevalence of LBP was higher using self-

reported (21.2%) than administrative data (10.2%), and agreement was low (kappa = 0.21). Administrative data had sensitivity 23.9% (95% CI 23.1-24.6), specificity 93.4% (95% CI 93.2-93.7), positive PV 50.4% (95% CI 49.1-51.7), and negative PV 82.0% (95% CI 81.7-82.3). Characteristics of LBP cases based on data sources differed in sex, health/behaviour characteristics, and allied health care utilization. Using health administrative data significantly underestimates the prevalence of LBP. This can lead to misclassification bias that is likely nondifferential in epidemiological studies.

Yasaitis, L. C., Berkman, L. F. et Chandra, A. (2015). "Comparison of self-reported and Medicare claims-identified acute myocardial infarction." Circulation **131**(17): 1477-1485; discussion 1485.

BACKGROUND: Cardiovascular disease is often studied through patient self-report and administrative data. However, these 2 sources provide different information, and few studies have compared them. METHODS AND RESULTS: We compared data from a longitudinal, nationally representative survey of older Americans with matched Medicare claims. Self-reported heart attack in the previous 2 years was compared with claims-identified acute myocardial infarction (AMI) and acute coronary syndrome. Among the 3.1% of respondents with self-reported heart attack, 32.8% had claims-identified AMI, 16.5% had non-AMI acute coronary syndrome, and 25.8% had other cardiac claims; 17.3% had no inpatient visits in the previous 2.5 years. Claims-identified AMIs were found in 1.4% of respondents; of these, 67.8% reported a heart attack. Self-reports were less likely among respondents >75 years of age (62.7% versus 74.6%; P=0.006), with less than high school education (61.6% versus 71.4%; P=0.015), with at least 1 limitation in activities of daily living (59.6% versus 74.7%; P=0.001), or below the 25th percentile of a word recall memory test (60.7% versus 71.3%; P=0.019). Both self-reported and claims-identified cardiac events were associated with increased mortality; the highest mortality was observed among those with claims-identified AMI who did not self-report (odds ratio, 2.8; 95% confidence interval, 1.5-5.1) and among those with self-reported heart attack and claims-identified AMI (odds ratio, 2.5; 95% confidence interval, 1.7-3.6) or non-AMI acute coronary syndrome (odds ratio, 2.7; 95% confidence interval, 1.8-4.1). CONCLUSIONS: There is considerable disagreement between self-reported and claims-identified events. Although self-reported heart attack may be inaccurate, it indicates increased risk of death, regardless of whether the self-report is confirmed by Medicare claims.

Zablotsky, B. et Black, L. I. (2019). "Concordance between survey reported childhood asthma and linked Medicaid administrative records." Journal of Asthma **56**(3): 285-295.

Objective: Agreement between administrative and survey data has been shown to vary by the condition of interest and there is limited research dedicated to parental report of asthma among children. The current study assesses the concordance between parent-reported asthma from the National Health Interview Survey (NHIS) with Medicaid administrative claims data among linkage eligible children from the NHIS. Methods: Medicaid Analytic eXtract (MAX) files from the Centers for Medicare & Medicaid Services (CMS) (years 2000-2005) were linked to participants of the NHIS (years 2001-2005). Concordance measures were calculated to assess overall agreement between a claims-based asthma diagnosis and a survey-based asthma diagnosis. Structural equation modeling was used to assess the association between demographic, service utilization, and co-occurring conditions factors and agreement. Results: Percent agreement between the two data sources was high (90%) with a prevalence-adjusted bias-adjusted kappa of 0.80 and Cohen's kappa of 0.55. Agreement varied by demographic characteristics, service utilization characteristics, and the presence of allergies and other health conditions. Structural equation modeling results found the presence of a series of co-occurring conditions, namely allergies, resulted in significantly lower agreement after controlling for demographics and service utilization. Conclusions: There was general agreement between asthma diagnoses reported in the NHIS when compared to medical claims. Discordance was greatest among children with co-occurring conditions.

Appariements entre les bases de données médico-administratives et les données d'enquêtes déclaratives

Antol, D. D., Hagan, A., Nguyen, H., et al. (2022). "Change in self-reported health: A signal for early intervention in a medicare population." Healthc (Amst) **10**(1): 100610.

> Background: Health plans and risk-bearing provider organizations seek information sources to inform proactive interventions for patients at risk of adverse health events. Interventions should take into account the strong relationship between social context and health. This retrospective cohort study of a Medicare Advantage population examined whether a change in self-reported health-related quality of life (HRQOL) signals a subsequent change in healthcare needs. Methods: A retrospective longitudinal analysis of administrative claims data was conducted for participants in a Medicare Advantage plan with prescription drug coverage (MAPD) who responded to 2 administrations of the Centers for Disease Control and Prevention 4-item Healthy Days survey within 6-18 months during 2015-2018. Changes in HRQOL, as measured by the Healthy Days instrument, were compared with changes in utilization and costs, which were considered to be a reflection of change in healthcare needs. Results: A total of 48,841 individuals met inclusion criteria. Declining HRQOL was followed by increases in utilization and costs. An adjusted analysis showed that every additional unhealthy day reported one year after baseline was accompanied by an $8 increase in monthly healthcare costs in the subsequent six months for the average patient. Conclusions: Declining HRQOL signaled subsequent increases in healthcare needs and utilization.

Avina-Galindo, A. M., Fazal, Z. A., Marozoff, S., et al. (2021). "Immunosuppression and COVID-19 infection in British Columbia: Protocol for a linkage study of population-based administrative and self-reported survey data." PLoS One **16**(11): e0259601.

> Introduction Cases of the novel coronavirus disease (COVID-19) continue to spread around the world even one year after the declaration of a global pandemic. Those with weakened immune systems, due to immunosuppressive medications or disease, may be at higher risk of COVID-19. This includes individuals with autoimmune diseases, cancer, transplants, and dialysis patients. Assessing the risk and outcomes of COVID-19 in this population has been challenging. While administrative databases provide data with minimal selection and recall bias, clinical and behavioral data is lacking. To address this, we are collecting self-reported survey data from a randomly selected subsample with and without COVID-19, which will be linked to administrative health data, to better quantify the risk of COVID-19 infection associated with immunosuppression. Methods and analysis Using administrative and laboratory data from British Columbia (BC), Canada, we established a population-based case-control study of all individuals who tested positive for SARS-CoV-2. Each case was matched to 40 randomly selected individuals from two control groups: individuals who tested negative for SARS-CoV-2 (i.e., negative controls) and untested individuals from the general population (i.e., untested controls). We will contact 1000 individuals from each group to complete a survey co-designed with patient partners. A conditional logistic regression model will adjust for potential confounders and effect modifiers. We will examine the odds of COVID-19 infection according to immunosuppressive medication or disease type. To adjust for relevant confounders and effect modifiers not available in administrative data, the survey will include questions on behavioural variables that influence probability of being tested, acquiring COVID-19, and experiencing severe outcomes. Ethics and dissemination This study has received approval from the University of British Columbia Clinical Research Ethics Board [H20-01914]. Findings will be disseminated through scientific conferences, open access peer-reviewed journals, COVID-19 research repositories and dissemination channels used by our patient partners.

Berete, F., Demarest, S., Charafeddine, R., et al. (2023). "Linking health survey data with health insurance data: methodology, challenges, opportunities and recommendations for public health research. An experience from the HISlink project in Belgium." Arch Public Health **81**(1): 198.

> In recent years, the linkage of survey data to health administrative data has increased. This offers new opportunities for research into the use of health services and public health. Building on the HISlink use

case, the linkage of Belgian Health Interview Survey (BHIS) data and Belgian Compulsory Health Insurance (BCHI) data, this paper provides an overview of the practical implementation of linking data, the outcomes in terms of a linked dataset and of the studies conducted as well as the lessons learned and recommendations for future links. Individual BHIS 2013 and 2018 data was linked to BCHI data using the national register number. The overall linkage rate was 92.3% and 94.2% for HISlink 2013 and HISlink 2018, respectively. Linked BHIS-BCHI data were used in validation studies (e.g. self-reported breast cancer screening; chronic diseases, polypharmacy), in policy-driven research (e.g., mediation effect of health literacy in the relationship between socioeconomic status and health related outcomes, and in longitudinal study (e.g. identifying predictors of nursing home admission among older BHIS participants). The linkage of both data sources combines their strengths but does not overcome all weaknesses.The availability of a national register number was an asset for HISlink. Policy-makers and researchers must take initiatives to find a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health. Researchers should be aware that the procedures necessary to implement a link may have an impact on the timeliness of their research. Although some aspects of HISlink are specific to the Belgian context, we believe that some lessons learned are useful in an international context, especially for other European Union member states that collect similar data.

Carrere, A. (2016). "Les enrichissements prévus pour l'enquête CARE-Ménages : mise en œuvre, apports et contraintes." Serie sources et Méthodes - Document de travail - Drees(56)

Les enquêtes CARE (Capacités, Aides et REssources des seniors) ont pour but de répondre à ces besoins d'information. Or, y répondre uniquement par la collecte auprès des personnes interrogées peut s'avérer inefficace et lourd. Les personnes, surtout âgées, peuvent ne pas communiquer à l'enquêteur l'ensemble des aides qu'elles reçoivent et des coûts qu'elles supportent, même avec l'aide d un tiers pour répondre à l'enquête, d'autant plus qu'elles peuvent bénéficier de remboursements ex-post. La question qui se pose est : comment disposer de suffisamment de données pour répondre de façon fiable aux besoins d'information des pouvoirs publics sans alourdir la charge des enquêtés ? Le code de bonnes pratiques de la statistique européenne rappelle, dans son principe 9, de ne pas soumettre une charge excessive pour les déclarants et, dans le cas où des sources administratives existent, l'article 7bis de la loi n° 51-711 du 7 juin 1951 prévoit la cession, à « des fins exclusives d'établissement de statistiques » des informations « recueillies dans le cadre de sa mission par une administration, une personne morale de droit public ou une personne morale de droit privé gérant un service public. » Par ailleurs, le Conseil national de l'information statistique (CNIS) fait état du souhait de disposer d'enquêtes en population générale sur la dépendance, mettant en relation les différents systèmes de gestion de la dépendance : organismes de sécurité sociale, conseils généraux, sources fiscales. Afin de respecter ce principe 9, et parce que les données administratives sont de qualité, il a été décidé de réaliser différents enrichissements par appariement des réponses à l'enquête à de telles données administratives. Le dispositif d'enquêtes CARE sera présenté dans une première partie. La deuxième partie s'intéressera aux objectifs des enrichissements envisagés. Enfin, seront présentées les contraintes et les difficultés des enrichissements dans une dernière partie.

Comino, E. J., Tran, D. T., Haas, M., et al. (2013). "Validating self-report of diabetes use by participants in the 45 and Up Study: a record linkage study." BMC Health Serv Res **13**: 481.

BACKGROUND: Prevalence studies usually depend on self-report of disease status in survey data or administrative data collections and may over- or under-estimate disease prevalence. The establishment of a linked data collection provided an opportunity to explore the accuracy and completeness of capture of information about diabetes in survey and administrative data collections. METHODS: Baseline questionnaire data at recruitment to the 45 and Up Study was obtained for 266,848 adults aged 45 years and over sampled from New South Wales, Australia in 2006-2009, and linked to administrative data about hospitalisation from the Admitted Patient Data Collection (APDC) for 2000-2009, claims for medical services (MBS) and pharmaceuticals (PBS) from Medicare Australia data for 2004-2009. Diabetes status was determined from response to a question 'Has a doctor EVER told you that you have diabetes' (n = 23,981) and augmented by examination of free text fields about diagnosis (n = 119) or use of insulin (n = 58). These data were used to identify the sub-group with type

1 diabetes. We explored the agreement between self-report of diabetes, identification of diabetes diagnostic codes in APDC data, claims for glycosylated haemoglobin (HbA1c) in MBS data, and claims for dispensed medication (oral hyperglycaemic agents and insulin) in PBS data. RESULTS: Most participants with diabetes were identified in APDC data if admitted to hospital (79.3%), in MBS data with at least one claim for HbA1c testing (84.7%; 73.4% if 2 tests claimed) or in PBS data through claim for diabetes medication (71.4%). Using these alternate data collections as an imperfect 'gold standard' we calculated sensitivities of 83.7% for APDC, 63.9% (80.5% for two tests) for MBS, and 96.6% for PBS data and specificities of 97.7%, 98.4% and 97.1% respectively. The lower sensitivity for HbA1c may reflect the use of this test to screen for diabetes suggesting that it is less useful in identifying people with diabetes without additional information. Kappa values were 0.80, 0.70 and 0.80 for APDC, MBS and PBS respectively reflecting the large population sample under consideration. Compared to APDC, there was poor agreement about identifying type 1 diabetes status. CONCLUSIONS: Self-report of diagnosis augmented with free text data indicating diabetes as a chronic condition and/or use of insulin among medications used was able to identify participants with diabetes with high sensitivity and specificity compared to available administrative data collections.

Druschke, D., Arnold, K., Heinrich, L., et al. (2020). "Individual-Level Linkage of Primary and Secondary Data from Three Sources for Comprehensive Analyses of Low Birthweight Effects." Gesundheitswesen **82**: S108-S116.

Aim of the study The linkage of primary and secondary data is becoming an increasingly popular approach in healthcare research, but involves some challenges for all involved parties, for example due to data protection requirements. The aim of this article is to systematically outline the methods used and experiences made during a cohort study in the field of pediatric health care research (EcoCare-PIn) that involved access to and linkage of three different data sources. Particular focus is placed on the necessary regulatory measures with regard to data access and data linkage as well as on data validation to ensure a correct linkage. Methods While complying with all relevant data protection requirements, the study realized an individual-level linkage of a) pseudonymized administrative health insurance data from a statutory health insurance on Saxon children born between 2007 and 2013, b) primary data collected via postal questionnaires from parents/caregivers and c) medical data from kindergarten- and school-entry-examinations of Saxon health authorities. The fundamental principle of the concept of data linkage was to strictly separate the sites of data collection and data analysis, which was realized through the involvement of a trust center. Results Challenges especially pertained to the extensive regulatory pre-requirements for data access as well as to data protection requirements while performing the study. Technical aspects and data validation also required a considerable share of attention and resources. A number of validation routines were applied to avoid incorrect data linkage and to ensure the high quality of the final dataset. Data validation included both plausibility checks within the primary data and consistency checks of information given in primary and secondary data. Conclusion The linkage of primary and secondary data on the individual level offers great opportunities for using the strengths of different data sources synergistically and overcoming some of their limitations. Statutory health insurance data and medical data from kindergarten- and school-entry-examinations of Saxon health authorities are examples of already existing data sources that can complement cost-consuming primary data collections by valuable data sets and open up opportunities for longitudinal analysis.

Hucteau, E., Noize, P., Pariente, A., et al. (2021). "ADL-dependent older adults were identified in medico-administrative databases." J Clin Epidemiol **139**: 297-306.

OBJECTIVE: We aimed to develop an algorithm for the identification of basic Activities of Daily Living (ADL)-dependency in health insurance databases. STUDY DESIGN AND SETTING: We used the AMI (Aging Multidisciplinary Investigation) population-based cohort including both individual face-to-face assessment of ADL-dependency and merged health insurance data. The health insurance factors associated with ADL-dependency were identified using a LASSO logistic regression model in 1000 bootstrap samples. An external validation on a 1/97 representative sample of the French Health Insurance general population of Affiliates has been performed. RESULTS: Among 995 participants of the AMI cohort aged >/= 65y, 114 (11.5%) were ADL-dependent according to neuropsychologists

individual assessments. The final algorithm developed included: age, sex, four drug classes (dopaminergic antiparkinson drugs, antidepressants, antidiabetic agents, lipid modifying agents), three type of medical devices (medical bed, patient lifter, incontinence equipment), four medical acts (GP's consultations at home, daily and non-daily nursing at home, transport by ambulance) and four long-term diseases (stroke, heart failure, coronary heart disease, Alzheimer and other dementia). Applying this algorithm, the estimated prevalence of ADL-dependency was 12.3% in AMI and 9.5% in the validation sample. CONCLUSION: This study proposes a useful algorithm to identify ADL-dependency in the health insurance data.

Kab, S. et Goldberg, M. (2026). "Survey-French national health data system (SNDS) linkage: A win-win methodology for longitudinal studies, algorithm validation, and real-world evidence." J Epidemiol Popul Health **74**(2): 203391.

BACKGROUND: Integrating granular personal and clinical data with large-scale administrative records is a frontier in modern public health. In France, linking national epidemiological surveys with the National Health Data System (SNDS)-one of the world's most exhaustive administrative databases-offers a transformative "win-win" methodology to overcome self-reporting biases and loss to follow-up. METHODS: This paper analyzes the architectural and methodological frameworks of data linkage to SNDS in France, distinguishing between deterministic linkage (via the National Identification Number) and probabilistic approaches. We examine major national integrations, including the prospective Constances cohort, cross-sectional surveys or clinical cohorts, and administrative cohorts like EDP-Santé. RESULTS: Linkage significantly enhances data utility by cross-referencing objective healthcare consumption with socio-economic, environmental, and behavioral health determinants. Beyond data enrichment, this synergy provides a robust methodological platform for the validation of identification algorithms, allowing researchers to calculate sensitivity and specificity against clinical "Gold Standards." We highlight how these linked datasets facilitate complex longitudinal studies on social health inequalities and care pathways that are unattainable through isolated sources. CONCLUSION: Survey-SNDS linkage is a "win-win" process that has become the foundational standard for high-impact research in France. By maximizing the utility of national data assets, this methodology provides a replicable model for global real-world evidence (RWE) generation and public health policy evaluation.

March, S. (2017). "Individual Data Linkage of Survey Data with Claims Data in Germany—An Overview Based on a Cohort Study." Ijerph **14**(12): 1-16.

Research based on health insurance data has a long tradition in Germany. By contrast, data linkage of survey data with such claims data is a relatively new field of research with high potential. Data linkage opens up new opportunities for analyses in the field of health services research and public health. Germany has comprehensive rules and regulations of data protection that have to be followed. Therefore, a written informed consent is needed for individual data linkage. Additionally, the health system is characterized by heterogeneity of health insurance. The lidA-living at work-study is a cohort study on work, age and health, which linked survey data with claims data of a large number of statutory health insurance data. All health insurance funds were contacted, of whom a written consent was given. This paper will give an overview of individual data linkage of survey data with German claims data on the example of the lidA-study results. The challenges and limitations of data linkage will be presented. Despite heterogeneity, such kind of studies is possible with a negligibly small influence of bias. The experience we gain in lidA will be shown and provide important insights for other studies focusing on data linkage.

McCormick, N., Reimer, K., Famouri, A., et al. (2017). "Filling the gaps in SARDs research: collection and linkage of administrative health data and self-reported survey data for a general population-based cohort of individuals with and without diagnoses of systemic autoimmune rheumatic disease (SARDs) from British Columbia, Canada." BMJ OPEN **7**(6).

Purpose Systemic autoimmune rheumatic diseases (SARDs) are a group of debilitating autoimmune diseases, including systemic lupus erythematosus and related disorders. Assessing the healthcare and

economic burden of SARDs has been challenging: while administrative databases can be used to determine healthcare utilisation and costs with minimal selection and recall bias, other health, sociodemographic and economic data have typically been sourced from highly selected, clinic-based cohorts. To address these gaps, we are collecting self-reported survey data from a general population-based cohort of individuals with and without SARDs and linking it to their longitudinal administrative health data. Participants Using administrative data from the province of British Columbia (BC), Canada, we established a population-based cohort of all BC adults receiving care for SARDs during 1996-2010 (n= 20 729) and non-SARD individuals randomly selected from the general population. BC Ministry of Health granted us contact information for 12 000 SARD and non-SARD individuals, who were recruited to complete the surveys by mail or online. Findings to date Four hundred individuals were initially invited to participate, with 135 (34%) consenting and 127 (94%) submitting the first survey (72% completed online). Sixty-three (49.6%) reported >= 1 SARD diagnosis. The non-SARDs group (n= 64) was 92% female with mean age 57.0 +/- 11.6 years. The SARDs group (n= 63) was 94% female with mean age 56.5 +/- 13.1 years. Forty-eight per cent of those with SARDs were current-or-former smokers (mean 10.6 +/- 16.2 pack-years), and 33% were overweight or obese (mean body mass index of 24.4 +/- 5.3). Future plans Health and productivity data collected from the surveys will be linked to participants' administrative health data from the years 1990-2013, allowing us to determine the healthcare and lost productivity costs of SARDs, and assess the impact of patient-reported variables on utilisation, costs, disability and clinical outcomes. Findings will be disseminated through scientific conferences and peer-reviewed journals.

Montaut, A., Calvet, L., Bouvier, G., et al. (2013). "L'appariement handicap-santé et données de l'assurance maladie." Série sources et Méthodes - document de travail - Drees(39)

[BDSP. Notice produite par MIN-SANTE IFl8R0x8. Diffusion soumise à autorisation]. L'enquête Handicap Santé permet de dresser un bilan de l'état de santé de la population. Elle s'intéresse en particulier aux personnes handicapées ou en situation de dépendance. Cette enquête comporte deux volets : en ménage (collecté en 2008) et en institution (collecté en 2009). L'enquête Handicap Santé a été appariée avec les données de l'Assurance Maladie, complétant ainsi l'enquête avec le recours aux soins et les dépenses de santé des personnes enquêtées. Ce document de travail présente l'appariement de ces deux sources : des difficultés juridiques aux traitements statistiques mis en oeuvre pour assurer la représentativité de l'échantillon apparié.

Ngo, P. J., Wade, S., Banks, E., et al. (2022). "Large-Scale Population-Based Surveys Linked to Administrative Health Databases as a Source of Data on Health Utilities in Australia." Value Health **25**(9): 1634-1643.

OBJECTIVES: Large-scale health surveys that contain quality-of-life instruments are a rich source of health utility data for health economic evaluations, especially when linked to routinely collected, administrative health databases. We derived health utility values for a wide range of health conditions using a large Australian cohort study linked to population-wide health databases. METHODS: Short-Form 6-Dimension utility values were calculated for 56 094 adults, aged 47+ years, in the New South Wales 45 and Up Study who completed the Social, Economic, and Environmental Factors survey (2010-2011). Mean utilities were summarized for major health conditions identified through self-report, hospital records, primary cancer notifications, and claims for government-subsidized prescription medicines and medical services. To identify unique associations between health conditions and utilities, beta regression was performed. Utility values were analyzed by time to death using linked death records. RESULTS: Mean Short-Form 6-Dimension utility was 0.810 (95% confidence interval [CI] 0.809-0.811), was age dependent, and was higher in men than women. Utilities for serious health conditions ranged from 0.685 (95% CI 0.652-0.718) for lung cancer to 0.800 (95% CI 0.787-0.812) for melanoma whereas disease-free respondents had a mean of 0.859 (95% CI 0.858-0.861). Most health conditions were independently associated with poorer quality of life. Utility values also declined by proximity to death where participants sampled 6 months before death had a mean score of 0.637 (95% CI 0.613-0.662). CONCLUSIONS: Our data offer a snapshot of the health status of an older Australian population and show that record linkage can enable comprehensive ascertainment of utility values for use in health economic modeling.

Ni, J. Y., Leong, A., Dasgupta, K., et al. (2017). "Correcting hazard ratio estimates for outcome misclassification using multiple imputation with internal validation data." Pharmacoepidemiol Drug Saf **26**(8): 925-934.

Objective Outcome misclassification may occur in observational studies using administrative databases. We evaluated a two-step multiple imputation approach based on complementary internal validation data obtained from two subsamples of study participants to reduce bias in hazard ratio (HR) estimates in Cox regressions. Methods We illustrated this approach using data from a surveyed sample of 6247 individuals in a study of statin-diabetes association in Quebec. We corrected diabetes status and onset assessed from health administrative data against self-reported diabetes and/or elevated fasting blood glucose (FBG) assessed in subsamples. The association between statin use and new onset diabetes was evaluated using administrative data and the corrected data. By simulation, we assessed the performance of this method varying the true HR, sensitivity, specificity, and the size of validation subsamples. Results The adjusted HR of new onset diabetes among statin users versus non-users was 1.61 (95% confidence interval: 1.09-2.38) using administrative data only, 1.49 (0.95-2.34) when diabetes status and onset were corrected based on self-report and undiagnosed diabetes (FBG >= 7 mmol/L), and 1.36 (0.92-2.01) when corrected for self-report and undiagnosed diabetes/impaired FBG (>= 6 mmol/L). In simulations, the multiple imputation approach yielded less biased HR estimates and appropriate coverage for both non-differential and differential misclassification. Large variations in the corrected HR estimates were observed using validation subsamples with low participation proportion. The bias correction was sometimes outweighed by the uncertainty introduced by the unknown time of event occurrence. Conclusion Multiple imputation is useful to correct for outcome misclassification in time-to-event analyses if complementary validation data are available from subsamples. Copyright (C) 2017 John Wiley & Sons, Ltd.

Olesen, S. C., Butterworth, P., Jacomb, P., et al. (2012). "Personal factors influence use of cervical cancer screening services: epidemiological survey and linked administrative data address the limitations of previous research." BMC Health Serv Res **12**: 34.

Background: National screening programs have reduced cervical cancer mortality; however participation in these programs varies according to women's personal and social characteristics. Research into these inequalities has been limited by reliance on self-reported service use data that is potentially biased, or administrative data that lacks personal detail. We address these limitations and extend existing research by examining rates and correlates of cervical screening in a large epidemiological survey with linked administrative data. Methods: The cross-sectional sample included 1685 women aged 44-48 and 64-68 years from the Australian Capital Territory and Queanbeyan, Australia. Relative risk was assessed by logistic regression models and summary Population Attributable Risk (PAR) was used to quantify the effect of inequalities on rates of cervical cancer screening. Results: Overall, 60.5% of women participated in screening over the two-year period recommended by Australian guidelines. Screening participation was associated with having children, moderate or high use of health services, employment, reported lifetime history of drug use, and better physical functioning. Conversely, rates of cervical screening were lower amongst women who were older, reliant on welfare, obese, current smokers, reported childhood sexual abuse, and those with anxiety symptoms. A summary PAR showed that effective targeting of women with readily observable risk-factors (no children, no partner, receiving income support payments, not working, obese, current smoker, anxiety, poor physical health, and low overall health service use) could potentially reduce overall non-participation in screening by 74%. Conclusions: This study illustrates a valuable method for investigating the personal determinants of health service use by combining representative survey data with linked administrative records. Reliable knowledge about the characteristics that predict uptake of cervical cancer screening services will inform targeted health promotion efforts.

Polin, K., Panteli, D. et Webb, E. (2023). "Health and Care Data : Approaches to data linkage for evidence-informed policy". Copenhague, O.M.S. Bureau régional de l'Europe
https://iris.who.int/bitstream/handle/10665/371097/9789289059466-eng.pdf

Evidence-based health policy requires good health services research, which in turn requires access to comprehensive high-quality data. With the digital transformation of healthcare, ever-more dynamic

landscapes of datasets, and the availability of 'big data', health services research increasingly relies on linking data within and outside of health for meaningful insights. Internationally, country approaches to data collection and processing for secondary research purposes (i.e., health systems research) vary. Data linkage requires common key variables to combine information from multiple sources. To understand these approaches and gain insights into good practices, exchange is vital. Based on 30 case studies across 13 high-income countries, this review provides an overview of existing practices in data linkage for health services research. The case studies include:large administrative datasets,centralized locations that merge existing datasets;databases which combine routinely collected data;patient-centric electronic platforms; and tools that facilitate research.

Raghunathan, T., Ghosh, K., Rosen, A., et al. (2021). "Combining Information From Multiple Data Sources to Assess Population Health." Journal of Survey Statistics and Methodology **9**(3): 598-625

Information about an extensive set of health conditions on a well-defined sample of subjects is essential for assessing population health, gauging the impact of various policies, modeling costs, and studying health disparities. Unfortunately, there is no single data source that provides accurate information about health conditions. We combine information from several administrative and survey data sets to obtain model-based dummy variables for 107 health conditions (diseases, preventive measures, and screening for diseases) for elderly (age 65 and older) subjects in the Medicare Current Beneficiary Survey (MCBS) over the fourteen-year period, 1999-2012. The MCBS has prevalence of diseases assessed based on Medicare claims and provides detailed information on all health conditions but is prone to underestimation bias. The National Health and Nutrition Examination Survey (NHANES), on the other hand, collects self-reports and physical/laboratory measures only for a subset of the 107 health conditions. Neither source provides complete information, but we use them together to derive model-based corrected dummy variables in MCBS for the full range of existing health conditions using a missing data and measurement error model framework. We create multiply imputed dummy variables and use them to construct the prevalence rate and trend estimates. The broader goal, however, is to use these corrected or modeled dummy variables for a multitude of policy analysis, cost modeling, and analysis of other relationships either using them as predictors or as outcome variables.

Rose, S., Xie, D. W., Streim, J. E., et al. (2016). "Identifying neuropsychiatric disorders in the Medicare Current Beneficiary Survey: the benefits of combining health survey and claims data." BMC Health Serv Res **16**.

Background: To address the impact of using multiple sources of data in the United States Medicare Current Beneficiary Survey (MCBS) compared to using only one source of data to identify those with neuropsychiatric diagnoses. Methods: Our data source was the 2010 MCBS with associated Medicare claims files (N = 14, 672 beneficiaries). The MCBS uses a stratified multistage probability sample design to select a nationally representative sample of Medicare beneficiaries. We excluded those participants in Medicare Health Maintenance Organizations (n = 3894) and performed a cross-sectional analysis. We classified neuropsychiatric conditions according to four broad categories: intellectual/developmental disorders, neurological conditions affecting the central nervous system (Neuro-CNS), dementia, and psychiatric conditions. To account for different baseline prevalence differences of the categories we calculated the relative increase in prevalence that occurred from adding information from claims in addition to the absolute increase to allow comparison among categories. Results: The estimated proportion of the sample with neuropsychiatric disorders increased to 50.0 (both sources) compared to 38.9 (health survey only) and 33.2 (claims only) with an overlap between sources of only 44.1 %. Augmenting health survey data with claims led to an increase in estimated percentage of intellectual/developmental disorders, psychiatric disorders, Neuro-CNS disorders and dementia of 1.3, 5.9, 11.5 and 3.8 respectively. In the community sample, the largest relative increases were seen for dementia (147.6 %) and Neuro-CNS disorders (87.4 %). With the exception of dementia, larger relative increases were seen in the facility sample with the greatest being for intellectual/developmental disorders (121.5 %) and Neuro-CNS disorders (93.8 %).Conclusions: The magnitude of potentially underestimated sample proportions using health survey only data varied strikingly according to the category of diagnosis and setting. Augmentation of survey data with claims appears essential particularly when attempting to estimate proportion of the sample

affected by conditions that cause cognitive impairment which may affect ability to self-report. Augmenting proxy survey data with claims data also appears to be essential when ascertaining proportion of the facility-dwelling sample affected by neuropsychiatric disorders.

Runte, R. (2018). "Predictors of institutionalization in people with dementia: a survey linked with administrative data." Aging Clin Exp Res **30**(1): 35-43.

BACKGROUND: For people with dementia, moving into a nursing home is usually considered at some point in time. Currently available information on predictors of institutionalization is often based on small sample sizes, not taking competing risks into account, and with inconclusive results for sex. AIMS: We aimed to carry out an analysis stratified by sex and using a competing risk approach. METHODS: We carried out an analysis of a survey linked with administrative data including 652 people with dementia, aged 60 years and older. The follow-up was up to 4.5 years. We used the cumulative incidence function for examining time until institutionalization and survival time and the sub-distribution hazard model for estimating hazard ratios. RESULTS: The participants were on average 81 years old, about 51% were female. At the end of the follow-up, 282 people had been institutionalized and 273 had died. The regression models show that the risk of institutionalization is higher in women than in men and when cared for by a care service in comparison to an informal caregiver. Inhibiting factors are Care Level (II, III) and positive evaluation of caregiving by caregivers. Stratified analysis by sex revealed that the risk of institutionalization in men is influenced by their relationship to their caregiver, in women by duration of care at baseline. DISCUSSION: Sex seems to play a role in predicting institutionalization. CONCLUSION: Future research should focus on stratified analysis by sex. Knowing the predictors of institutionalization for men and women could influence long-term care management remarkably.

Schüssler-Fiorenza Rose, S. M., Xie, D., Streim, J. E., et al. (2016). "Identifying neuropsychiatric disorders in the Medicare Current Beneficiary Survey: the benefits of combining health survey and claims data." BMC Health Serv Res **16**(1): 537.

BACKGROUND: To address the impact of using multiple sources of data in the United States Medicare Current Beneficiary Survey (MCBS) compared to using only one source of data to identify those with neuropsychiatric diagnoses. METHODS: Our data source was the 2010 MCBS with associated Medicare claims files (N = 14, 672 beneficiaries). The MCBS uses a stratified multistage probability sample design to select a nationally representative sample of Medicare beneficiaries. We excluded those participants in Medicare Health Maintenance Organizations (n = 3894) and performed a cross-sectional analysis. We classified neuropsychiatric conditions according to four broad categories: intellectual/developmental disorders, neurological conditions affecting the central nervous system (Neuro-CNS), dementia, and psychiatric conditions. To account for different baseline prevalence differences of the categories we calculated the relative increase in prevalence that occurred from adding information from claims in addition to the absolute increase to allow comparison among categories. RESULTS: The estimated proportion of the sample with neuropsychiatric disorders increased to 50.0 (both sources) compared to 38.9 (health survey only) and 33.2 (claims only) with an overlap between sources of only 44.1 %. Augmenting health survey data with claims led to an increase in estimated percentage of intellectual/developmental disorders, psychiatric disorders, Neuro-CNS disorders and dementia of 1.3, 5.9, 11.5 and 3.8 respectively. In the community sample, the largest relative increases were seen for dementia (147.6 %) and Neuro-CNS disorders (87.4 %). With the exception of dementia, larger relative increases were seen in the facility sample with the greatest being for intellectual/developmental disorders (121.5 %) and Neuro-CNS disorders (93.8 %). CONCLUSIONS: The magnitude of potentially underestimated sample proportions using health survey only data varied strikingly according to the category of diagnosis and setting. Augmentation of survey data with claims appears essential particularly when attempting to estimate proportion of the sample affected by conditions that cause cognitive impairment which may affect ability to self-report. Augmenting proxy survey data with claims data also appears to be essential when ascertaining proportion of the facility-dwelling sample affected by neuropsychiatric disorders.

Sibley, L. M., Moineddin, R., Agha, M. M., et al. (2010). "Risk Adjustment Using Administrative Data-Based and Survey-Derived Methods for Explaining Physician Utilization." Medical Care **48**(2): 175-182.

[Objectives: The objective of this study was to evaluate an administrative data-based risk adjustment method for predicting physician utilization and the contribution of survey-derived indicators of health status. The results of this study will support the use of administrative data for planning, reimbursement, and assessing equity of physician utilization. Methods: The Ontario portion of the 2000–2001 Canadian Community Health Survey was linked with administrative physician claims data from 2002–2003 and 2003–2004. Explanatory models of family physician (FP) and specialist physician (SP) utilization were run using demographic information and The Johns Hopkins University Adjusted Clinical Groups (ACG) Case-mix System. Survey-based measures of health status were then added to the models. The coefficient of determination, R, indicated the models' explanatory power. Results: The study sample consisted of 25,558 individuals aged 20 to 79 years representing approximately 7.8 million people. Over the 2 years of study period, 82.5% of the study population had a FP visit with a median of 6 visits and 53.2% had a SP visit with a median of 1 visit. The R values based on administrative data alone were 33% and 21% for the frequency of FP and SP visits and 16% and 35% for having one or more visit to an FPs and SPs, respectively. The addition of the survey-based measures to the administrative data-based models produced less than a 2% increase in explanatory power for any outcome. Conclusion: Administrative data-based measures of morbidity burden are valid and useful indicators of future physician utilization. The survey-derived measures used in this study did not contribute significantly to models on the basis of administrative data-based measures. These findings support the future use of administrative data-based data and Adjusted Clinical Groups for planning, reimbursement, and research.

Van der Heyden, J., Van Oyen, H., Berger, N., et al. (2015). "Activity limitations predict health care expenditures in the general population in Belgium." BMC Public Health **15**.

Background: Disability and chronic conditions both have an impact on health expenditures and although they are conceptually related, they present different dimensions of ill-health. Recent concepts of disability combine a biological understanding of impairment with the social dimension of activity limitation and resulted in the development of the Global Activity Limitation Indicator (GALI). This paper reports on the predictive value of the GALI on health care expenditures in relation to the presence of chronic conditions. Methods: Data from the Belgian Health Interview Survey 2008 were linked with data from the compulsory national health insurance (n = 7,286). The effect of activity limitation on health care expenditures was assessed via cost ratios from multivariate linear regression models. To study the factors contributing to the difference in health expenditure between persons with and without activity limitations, the Blinder-Oaxaca decomposition method was used. Results: Activity limitations are a strong determinant of health care expenditures. People with severe activity limitations (5.1%) accounted for 16.9% of the total health expenditure, whereas those without activity limitations (79.0%), were responsible for 51.5% of the total health expenditure. These observed differences in health care expenditures can to some extent be explained by chronic conditions, but activity limitations also contribute substantially to higher health care expenditures in the absence of chronic conditions (cost ratio 2.46; 95% CI 1.74-3.48 for moderate and 4.45; 95% CI 2.47-8.02 for severe activity limitations). The association between activity limitation and health care expenditures is stronger for reimbursed health care costs than for out-of-pocket payments. Conclusion: In the absence of chronic conditions, activity limitations appear to be an important determinant of health care expenditures. To make projections on health care expenditures, routine data on activity limitations are essential and complementary to data on chronic conditions.